AFRL-RH-WP-TR-2012-0199

SPEECH AND LANGUAGE AND LANGUAGE TRANSLATION (SALT)

Katherine M. Young
N-Space Analysis, LLC
305 Winding Trail
Xenia, OH 45385

Jeremy N. Gwinnup
Brian M. Ore
Michael R. Hutt
Stephen A. Thorn
David M. Hoeferlin
Jeff Cress
SRA International
5000 Springfield Street, Suite 200
Dayton, OH 45431

Final Report
December 2012

Approved for public release; distribution is unlimited.

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING
HUMAN EFFECTIVENESS DIRECTORATE
HUMAN-CENTERED ISR DIVISION
WRIGHT-PATTERSON AFB OH 45433
UNIITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2013-0199 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//SIGNED//

TIMOTHY R. ANDERSON, PhD.
Work Unit Manager
Human Trust and Interaction Branch

//SIGNED//

LOUISE A. CARTER, PhD.
Human-Centered ISR Division
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government’s approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* 15-12-2012	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* 13 August 2009 – 28 February 2013

4. TITLE AND SUBTITLE
Speech and Language and Language Translation (SALT)

5a. CONTRACT NUMBER
FA8650-09-D-6939-0013

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
*Katherine M. Young
**Jeremy N. Gwinnup
**Brian M. Ore
**Michael R. Hutt
**Stephen A. Thorn
**David M. Hoeferlin
**Jeff Cress

5d. PROJECT NUMBER

5e. TASK NUMBER
0013

5f. WORK UNIT NUMBER
H06D (7184X10C)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

*N-Space Analysis LLC
305 Winding Trail
Xenia, OH 45385

**SRA International, Inc.
5000 Springfield Street, Suite 200
Dayton, OH 45431

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Human-Centered ISR Division
Human Trust and Interaction Branch
Wright-Patterson Air Force Base, OH 45433

10. SPONSOR/MONITOR'S ACRONYM(S)
711 HPW/RHXS

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RH-WP-TR-2012-0199

12. DISTRIBUTION AVAILABILITY STATEMENT
Distribution A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
88ABW-2013-2896, cleared 18 June 2013

14. ABSTRACT
This final report provides research results in the development and utility of automatic speech recognition (ASR), machine translation (MT), natural language processing (NLP), speech synthesis (SS) and other speech and language processing technologies.

15. SUBJECT TERMS
automatic speech recognition (ASR), machine translation (MT), natural language processing (NLP), and speech synthesis (SS).

16. SECURITY CLASSIFICATION OF: Unclassified			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Timothy R. Anderson
U	U	U	SAR	127	**19b. TELEPONE NUMBER (*Include area code)*** N/A

THIS PAGE LEFT INTENTIONALLY BLANK.

TABLE OF CONTENTS

Section **Page**

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

Table **Page**

SUMMARY

This document provides a summary of work completed by SRA, for the period 20 August 2009 to 28 February 2013 under contract FA8650-09-D-6939-0013.

Air Force (AF) and Department of Defense (DoD) personnel are called upon to operate all over the world with the Global War on Terror, humanitarian relief operations, various coalition operations, foreign internal defense, etc. With its global reach and responsibilities, the DoD needs to monitor and understand ongoing situations, to anticipate new situations that will require responses, and to influence the outcomes. Much of the information needed to effectively understand, anticipate, and influence these situations and to operate in them is found in foreign language speech and text; however, there is a critical lack of linguists to understand and/or translate this material. To address the linguist shortfall and the sheer volume of potentially applicable foreign language material for military applications, the Human Language Technology (HLT) group of 711 HPW/RHXS is investigating the development and utility of automatic speech recognition (ASR), machine translation (MT), natural language processing (NLP), speech synthesis (SS) and other speech and language processing technologies.

The overall purpose of this work is to support research being conducted within the 711 HPW/RHXS's Speech and Communication Research, Engineering, Analysis, and Modeling (SCREAM) laboratory. Work in the SCREAM laboratory includes the development of capabilities for speech recognition of foreign languages and speech-to-speech translation.

The focus of this task will be on ASR, MT, and SS, especially methods for rapidly developing spoken language translation systems (SLTS) and text translation system (TTS) in new domains and languages, especially Less Commonly Taught Languages (LCTL) of interest to current and future military operations. In particular, this subtask seeks to develop MT and SS algorithms that are frugal in their requirements for human-transcribed training data.

1.0 INTRODUCTION

This document provides a summary of work completed by SRA International for the period 20 August 2009 to 28 February 2013 under contract FA8650-09-D-6939-00014.

Section 2 describes experiments and accomplishments in Automatic Speech Recognition, Machine Translation (MT), Speech Synthesis (SS), Spoken Language Translation Systems (SLTS) / Text Translation Systems (TTS), Laboratory Corpora Support, Named Entity (NE) Detection Benefits, Recognition and Translation Detection Performance Assessment, and English to Urdu Translation, as well as system administration tasks completed that support the research environment.

Section 3 summarizes conclusions drawn from the experiments and makes recommendations for future efforts.

2.0 EXPERIMENTS AND ACCOMPLISHMENTS

This section discusses experiments and accomplishments. Section 2.1 covers ASR, Section 2.2 covers MT, Section 2.3 covers SS, Section 2.4 covers SLTS / TTS, Section 2.5 covers Laboratory Corpora Support, Section 2.6 covers NE Detection Benefits, Section 2.7 covers Recognition and Translation Detection Performance Assessment, Section 2.8 covers English to Urdu Translation and Section 2.8.6 covers Laboratory System Administration Support.

Under the Information Operations Cyber Exploitation Research (ICER) contract, work on SMT, ASR, MT, SLTS, TTS, NE, and related technologies are included in multiple task orders that support the SCREAM Laboratory. In some cases these task orders may focus on different languages of interest, however many of the software utilities, algorithms, and procedures many be applicable to many languages and tasks. Some of the significant experiments and accomplishments described in this report may also appear in reports for other task orders as the work was split between multiple task orders.

2.1 Automatic Speech Recognition

This section discusses the ASR experiments that were evaluated. Section 0 discusses how the Sphinx-4 speech recognition engine was modified to apply class-based feature transforms and correctly mark Language Model (LM) scores in lattices. Recognition results obtained with Sphinx-4 are presented for Dari, English, Mandarin, and Pashto. Section 0 describes how an Arabic ASR system was developed on the Topic Detection and Tracking (TDT4) corpus. Section 0 describes the ASR system components that were developed for the International Workshop on Spoken Language Translation (IWSLT) 2011 evaluation. Finally, Section 0 summarizes the ASR experiments and Section 0 provides recommendations for future work.

2.1.1. Sphinx-4 Modifications and Decoding Experiments

This section discusses two modifications that were made to the Sphinx-4 speech recognition engine [1] and presents results obtained on Dari, English, Mandarin, and Pashto. First, support was added for applying class-based feature transforms. Feature transforms are used to reduce the mismatch between an Acoustic Model (AM) and a set of feature vectors. Given a feature vector o, a transformation matrix A, and a bias vector b, a transformed feature vector can be calculated as $Ao + b$. In the simplest case, we can apply a global feature transform whereby one transformation matrix and bias vector is applied to all features. Alternatively, class-based transforms group different Gaussian components into classes and apply a different transformation to each class. This was implemented in Sphinx-4 by allowing the user to specify an additional model file that associates each Gaussian with a class, and a feature transform file generated using the Hidden Markov Model (HMM) ToolKit (HTK) [2]. No support was added for estimating the classes or the feature transforms because this is already provided in HTK.

Second, an error was discovered in the way that Sphinx-4 marks the LM scores in lattices. Lattices represent multiple hypothesized word sequences using nodes to represent words and arcs to define the allowable word sequences. Each arc includes an AM score and an LM score. Figure 1 shows an example of a lattice. Because of the way that alternative hypotheses are stored in Sphinx-4, the same LM score was applied to all arcs that end in the same word. For example, in Figure 1 the LM probability associated with the arc connecting *a* to *didn't* is different from the LM probability associated with the arc connecting *Noah* to *didn't*; however, both of these arcs were labeled with the same LM probability in Sphinx-4. This was corrected

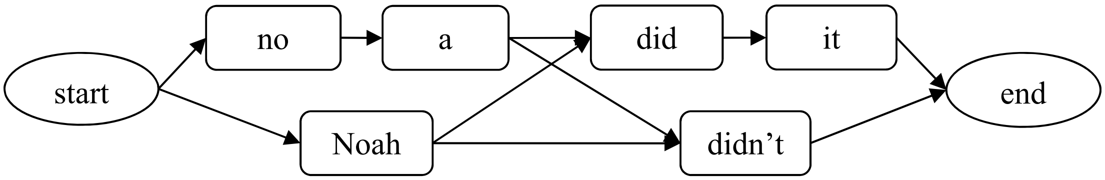


Figure 1: Example Word Lattice

Each word is represented by a node and the arcs represent the possible paths.

by modifying the alternate hypothesis manager to store current tokens instead of predecessor tokens.

Lastly, Sphinx-4 was evaluated on Dari, English, Mandarin, and Pashto. The AMs were developed using HTK and the LMs were estimated using the Stanford Research Institute Language Modeling (SRILM) Toolkit [3]. The systems were evaluated on the following corpora: Dari and Pashto on the Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) corpus; English on seven conference presentations downloaded from Technology, Entertainment, and Design (TED);[1] and Mandarin on the HUB4 Broadcast News Speech corpus [4]. This experiment investigated the effects of two different pruning parameters: the maximum number of active HMM states and the maximum number of active words. Table 1 shows the shows the real-time factor and error rate obtained for each language and parameter setting. The error rate for Dari, English, and Pashto is Word Error Rate (WER), and the error rate for Mandarin is Character Error Rate (CER).

2.1.2. Arabic ASR System

An Arabic speech recognition system was developed on the TDT4 Multilingual Broadcast News Speech corpus [5]. Whereas ASR corpora are typically time aligned at the sentence or phrase level, the TDT4 corpus is time aligned at the story level. This can create a problem when performing discriminative training because the lattices become very large due to the long file length. The following procedure was used to create shorter segments for this corpus. First, an HMM system was trained on the data using Maximum Likelihood (ML) estimation. Next, word alignments were automatically generated by forced alignment and each story was segmented into shorter utterances by iteratively splitting on long pauses. This procedure reduced the average file length from 70 seconds to 10 seconds. In addition, 32 minutes of English speech was identified and removed.

Wideband and narrowband AMs were trained on the TDT4 corpus using HTK. The bandwidth of each segment was automatically classified using a Gaussian Mixture Model (GMM) based bandwidth detector. Whereas the wideband models were trained on only the speech data that was classified as wideband, the narrowband models were estimated on all of the data by limiting the filterbank analysis from 125–3800 Hz. Phonemes were modeled using state clustered across word triphones, and the HMMs were discriminatively trained using the Minimum Phone Error (MPE) criterion. The final HMM sets included 4500 shared states with an average of 20 mixtures per state. The feature set for each AM consisted of 12 Mel-Frequency Cepstral

[1] Available at: http://www.ted.com

Table 1: Sphinx-4 Results on Dari, English, Mandarin, and Pashto

The error rate for Dari, English, and Pashto is WER, and the error rate for Mandarin is CER.

Language	HMM Pruning	Word Pruning	Real-Time factor	Error Rate
Dari	25k	50	1.39	41.6
	10k	25	0.93	42.1
English	25k	50	1.81	31.1
	10k	25	1.16	32.2
Mandarin	25k	50	2.11	19.3
	10k	25	0.89	19.6
Pashto	25k	50	0.88	37.7
	10k	25	0.67	38.1

Coefficients (MFCCs), plus energy, with mean normalization applied on a per utterance basis. Delta, acceleration, and third differential coefficients were appended to form a 52 dimensional feature vector, and Heteroscedastic Linear Discriminant Analysis (HLDA) was applied to reduce the feature dimension to 39.

A trigram LM was estimated on the training transcripts using the SRILM Toolkit. These models were evaluated on four shows from the TDT4 corpus that were excluded from the training data. Decoding was performed in a single pass using the HTK Large Vocabulary Continuous Speech Recognizer (LVCSR) HDecode. Using bandwidth dependent AMs yielded a 29.5% WER and decoding all segments using the narrowband AMs yielded a 29.3% WER. Note that the narrowband HMMs were trained with more speech data that the wideband HMMs because approximately 30% of the training utterances were classified as narrowband.

2.1.3. IWLST 2011 ASR System

This section discusses the ASR system components that were developed for the IWSLT 2011 evaluation.[2] AMs were developed on TED Talks, lattice rescoring was applied using 4-gram and 5-gram LMs, and system combination was investigated.

Acoustic data for training AMs were harvested from TED. A total of 826 TED Talk videos and transcripts were downloaded from the internet, and FFmpeg[3] was used to extract 16 kHz audio from each video file. Note that the transcripts provided by TED are not always an exact match to what was spoken and include only coarse time alignments. The following procedure was used to remove incorrectly transcribed sections and generate utterance level time alignments for the TED data. First, the time marks from the transcripts were used to remove long segments of untranscribed audio from each talk. Since only approximate time alignments are included in the transcripts, each speech segment was first padded by five seconds, and then overlapping sections were joined to form a single segment. A total of 1251 speech segments were extracted from the 826 TED Talks.

Next, word alignments were automatically generated for each speech segment using an HTK system trained on the HUB4 English Broadcast News corpora [6, 7]. Phonemes were modeled using state clustered across word triphones, and the HMMs were discriminatively trained using

[2] http://iwslt2011.org

[3] Available at: http://ffmpeg.org

the MPE criterion. The final HMM set included 4500 shared states with an average of 24 mixtures per state. The feature set included 12 Perceptual Linear Prediction (PLP) coefficients, plus the zeroth coefficient, with mean normalization applied on per utterance basis. Delta, acceleration, and third differential coefficients were appended to form a 52 dimensional feature vector, and HLDA was applied to reduce the feature dimension to 39. A total of 1224 segments from 807 different TED Talks were successfully aligned. The word alignments were used to further split each segment into utterances appropriate for training a speech recognition system. Utterance boundaries were first defined by any pause greater than $L = 0.75$ seconds. Next, L was decreased by 0.1 and all utterances longer than 20 seconds were reprocessed. This procedure was repeated until all segments were shorter than 20 seconds or $L = 0.15$. This yielded 85504 utterances.

Finally, closed caption filtering [8] was applied to sequester utterances from the TED data that may include transcription errors. Each utterance was decoded using HDecode with the HUB4 HMMs and a trigram LM that was estimated on the transcripts for that TED Talk. All LMs discussed in this section were trained using the SRILM Toolkit. The recognizer outputs were compared to the transcripts and the WER was calculated for each utterance. The WER across all files was 7.1%. A total of four different partitions were created by selecting utterances with WERs less than 20%, 30%, 50%, 100%, and an HMM system was trained on each partition using HTK. The HMMs used the sample model typology and feature set as the HUB4 HMMs, except that the HMMs included 4500–7500 shared states with an average of 20–28 mixtures per state and discriminative training was not applied to reduce development time. The IWSLT dev2010 partition was decoded using each HMM set and a trigram LM that was estimated from the TED text data provided by IWSLT. Table 2 shows the WERs obtained.

The final training partition was created by selecting utterances with a WER less than 20%, and the HMMs included 6000 shared states with an average of 28 mixtures per state. This system was re-evaluated on the IWSLT dev2010 partition using a two pass recognition strategy with Constrained Maximum Likelihood Linear Regression (CMLLR) transforms in the second pass. This system yielded a 26.9% WER. Applying discriminative training reduced the WER to 23.6%, and applying Speaker Adaptive Training (SAT) reduced the WER to 22.6%. Decoding was also performed using an interpolated trigram LM estimated from the following IWSLT 2011 data sets: TED, Europarl, News Commentary, and News 2007–2011. This system yielded a 19.5% WER.

The previous experiments were evaluated using automatically derived segments for the dev2010 partition. An additional experiment was performed using the time marks provided in the reference files to derive the segments. This reduced the WER to 18.5%. Next, the lattices output from the second pass recognizer were rescored using 4-gram and 5-gram interpolated LMs. These LMs were trained on the same data set as the trigram LM used in the recognizer. Rescoring with the 4-gram LM yielded a 17.8% WER and rescoring with the 5-gram LM yielded an 18.0% WER. Based on these results, the ASR system for the IWSLT 2011 evaluation utilized 4-gram rescoring. This system is referred to as *TED-IWSLT11* in the remainder of this section.

Table 2: WERs on IWSLT Dev2010 Obtained Using AMS Trained on TED Data

Closed caption filtering was applied using four different WER selection cutoffs. HMMs were trained with 4500–7000 shared states and an average of 20–28 mixtures per state.

WER Cutoff	Hours	Shared States	Average Mixtures	WER
20%	164	4500	20	30.2
		4500	24	29.6
		4500	28	29.1
		6000	20	29.4
		6000	24	29.0
		6000	28	28.5
		7500	20	29.0
		7500	24	28.8
		7500	28	28.4
30%	172	4500	20	30.4
		4500	24	29.5
		4500	28	29.0
		6000	20	29.5
		6000	24	29.0
		6000	28	28.5
		7500	20	29.1
		7500	24	28.7
		7500	28	28.4
50%	177	6000	20	29.6
		6000	24	28.9
		6000	28	28.7
100%	180	6000	20	29.8
		6000	24	29.4
		6000	28	28.9

The final experiment attempted to combine the output from three different ASR systems. The first system was the TED-IWSLT11 system. The second system used the same LMs and decoding strategy as the TED-IWSLT11 system, except that the AMs were trained on the HUB4 English Broadcast News Speech corpus. The final system was developed on TED data using the Kaldi Speech Recognition Toolkit [9]. The WERs of the individual systems were 17.8%, 22.6%, and 27.7%, respectively. Combining the outputs using the SRover software[4] from the University of Brno yielded an 18.6% WER, and combining the outputs using the rover software[5] from the National Institute of Standards and Technology (NIST) yielded a 19.2% WER.

[4] Available at: http://speech.fit.vutbr.cz/software/hmm-toolkit-stk

[5] Available at: http://www.itl.nist.gov/iad/mig/tools

2.1.4. Summary

The Sphinx-4 speech recognition engine was modified to apply class-based feature transforms and correctly mark LM scores in lattices. The recognizer performance was compared using two different pruning parameters on Dari, English, Mandarin, and Pashto. The best error rates obtained for each language were as follows: 41.6% WER on Dari, 31.1% WER on English, 19.3% CER on Mandarin, and 37.7% WER on Pashto.

An Arabic ASR system was developed on the TDT4 corpus. The story level transcripts were automatically segmented into shorter utterances so that discriminative training could be applied. Narrowband and wideband AMs were developed using HTK, and a trigram LM was estimated using the SRILM Toolkit. Decoding the development set using bandwidth dependent AMs yielded a 29.5% WER, and decoding all segments with the narrowband AMs yielded a 29.3% WER.

An English ASR system was developed for the IWSLT 2011 evaluation. Acoustic training data were harvested from TED and time aligned transcripts were automatically created using closed caption filtering. AMs were developed using HTK and evaluated on the IWSLT dev2010 partition. An improvement in system performance was obtained by discriminatively training the HMMs, applying SAT, decoding with an interpolated LM, and rescoring with a 4-gram LM. The best WER was 17.8%.

2.1.5. Recommendations for Future Work

Compared to recognition lattices generated by HDecode, Sphinx-4 lattices do not typically include as many hypothesized word sequences. It might be beneficial to investigate alternative algorithms for creating the lattices.

The Arabic ASR system did not apply SAT because the speaker identities were not known. One possible workaround would be to automatically cluster the segments. Another potential area for improvement would be using additional text data to estimate the LM.

No improvement in WER was obtained on the IWSLT dev2010 partition by combining the outputs from three different ASR systems. This may be because the difference in WER between the best and worst system is 9.9% absolute. An improvement in system performance might be obtained by combining systems with more comparable WERs or using a different method for system combination.

2.2 Machine Translation

Statistical MT is one of the main research efforts of the 711HPW/RHXS SCREAM Laboratory. Research included the development and extension of the Massachusetts Institute of Technology Lincoln Laboratory (MIT/LL) MIT/LL-AFRL SMT System (SMT2).

2.2.1. SMT2 Improvements

The SCREAM Lab works with MIT/LL on the SMT2 SMT system. In the course of various experiments and evaluations, extensions to the system are necessary to improve performance and implement additional functionality.

Word Aligners: Word alignment is an important step in the SMT process that occurs after sentence alignment. After sentence alignment identifies sentences that are translations of one another, the process is repeated at the word level. Word alignment may result in one-to-one or one-to-many mappings of words from the source to the target language.

Several word alignment programs were tested to see if they could improve the overall word alignment of the SMT2 system. The Posterior Constrained Toolkit (PostCAT) word aligner performance is less than that of GIZA, the standard word aligner. However, combining GIZA and PostCAT provides higher scores than GIZA alone.

The NILE word aligner was also tested, however not all portions of the alignment process have been implemented to run in SMT2. The NILE training process still needs to be run before starting an experiment.

Word Alignment Filtering: A library and series of tools named A3Metric were created to examine and modify results of word alignment. These utilities allow link editing/removal, rescoring based on different algorithms, and various alignment conversions.

A3PartitionFilter allows inclusion of out-of-domain data when performing word alignment. Once alignment is complete, the out-of-domain alignments are removed before creating the phrase table. Including the additional out-of-domain data improves quality of all the word alignments, but would make the phrase table unnecessarily large if the additional data were retained.

Additional experiments were performed using stemmed surface forms in the word alignment step and then restoring the original form. Results show a small improvement but further experimentation is necessary.

Input Adapters: An adapter was created to use compressed data directly from the Gigaword corpus in LM training. This adapter is able to uncompress the data and parse the file to retrieve the text data. Working directly with the input data saves disk space since an uncompressed copy is no longer needed. This approach can be applied to other formats in the future

One approach to improve the quality of word alignments in the training phase is to include out-of-domain data in order to provide more examples of various word correspondences. A method was constructed to place a boundary between the in-domain and out-of-domain data and align as usual, taking advantage of the additional data. Once word alignment is complete, the out-of-domain data can be discarded before creating the phrase table.

Language Model: Experiments were performed with the creation of an additional LM where certain classes of words were replaced with tags, (e.g. replacing a proper name with NAME.) These new LMs were used in conjunction with a special rescorer during optimization to leverage the additional information created in these models.

Additional experiments were run where extra LMs were created with stemmed forms of word and another where the surface form of the words was truncated after four characters. Training of systems was completed, but the optimized experiments have yet to be run for each of the two new LMs.

Rescorers: To compliment the work described in the above section, a rescorer was implemented for use during the optimization process. The TagRepLMRescorer takes the best entries from the list of possible translations, applies the same tag-matching algorithm to that entry and then scores it against the previously generated TagRep LM. When combined with other rescorers, a modest improvement over the baseline experiment was achieved.

Future work includes running similar experiments against stemmed LMs and Truncated-surface-form LMs.

Phrase Table Filters: Work was performed to wrap phrase table filtering scripts distributed with more recent versions of the Moses codebase. TMCombine and FillUp components were created to execute these scripts from within SMT2.

Confusion Net and Lattice Inputs to SMT2: One of the ways SMT2 was extended was the addition of new input data types to handle paraphrasing. Paraphrased sentences can be represented in different directed graph structures. Confusion networks represent possible different words as stacked nodes with the arc assigned to a probability weight representing how likely a given choice is. Lattices are similar, but each different phrase can be represented without having to connect to the other choices. The arcs still represent word probabilities.

Since lattices and confusion networks are used as input during the optimization and rescoring processes, changes had to be made to properly ingest the new input types. While lattices are represented on a single line with a parenthesis-delimited form, confusion networks span multiple lines. Since we need to maintain a one-to-one correspondence per line in SMT2, the confusion networks are flattened by replacing line breaks with a marker string (“&&&”). This string was chosen since it wouldn’t normally appear in data. The confusion networks are unpacked once the data has been read into the decoding portion of SMT2. Additional weight parameters had to be exposed for Moses to properly work. This weight parameter is responsible for arc-transition properties (iweight in Moses documentation). Another change necessary is to increase the maximum-phrase length to a large number, generally over 10,000. When running with confusion networks or lattices, this parameter refers to the maximum path length over all paths instead of just the length of the sentence.

Systems were run on the TED Talk English-to-French data set from IWSLT.

Paraphrased confusion networks and lattices were produced with differing numbers of possible paths (typically 3 and 10 alternative versions), unfortunately these paraphrased systems performed worse than the baseline system, but there are additional approaches to explore.

2.2.2. Character-Level Processing

MT is generally improved by reducing variation at the character level. This section reports on the reduction of character-level variation via spelling normalization and tokenization, in Dari, Pashto, Urdu, and French.

Spelling Normalization: Spelling normalization improves processing by conflating alternate forms of the same word. Control characters are also handled at this stage. This is accomplished in a separate step, since there are times when control characters should be maintained. In most cases, normalization should be applied to both the data and the lexicon: Spelling normalization will collapse lexicon entries with variant spellings; then, for words that are entered in the lexicon with only one spelling, normalization of the data will allow us to find the lexicon entry for either variant.

Dari and Pashto Spelling Normalization: The SCREAM lab normalization program for Arabic and Urdu, uninorm.pl, was adapted to also handle Dari and Pashto. The uninorm program maps Arabic presentation forms to their basic forms and removes diacritics; subroutines then perform the language-specific normalization.

Dari and Pashto both exhibit variations in the vowels forms; for example, the Dari word /nym/ 'half' can be spelled with either ي 064A or ی 06CC. The decision to normalize such forms depends on their relative frequency within the data set.

The Dari and Pashto data are compiled from the Sada-e-Azadi news website. No presentation forms have been noted in this Dari and Pashto data. There are some control characters in the newer Sada-e-Azadi data.

The following conversions are applied to the Pashto data:

Arabic digits 0660-0669 > Farsi digits 06F0-06F9
Farsi yeh 06CC ی > Arabic yeh 064A ي

Spelling variation is also an issue for the lexicon. There are entries that differ only in the choice of Farsi yeh vs. Arabic yeh; for example, سفير "ambassador" and سفیر "ambassador", with 064A ي and 06CC ی , respectively. Typically, the same spelling normalization should be applied to both the data and the lexicon. However, the Pashto data and the Linguistic Data Consortium (LDC) Pashto Lexicon differ in the amount of spelling variation.

The Pashto text primarily represents /k/ with the Farsi kaf ک, with a few instances of Arabic kaf ك; this suggests the presence of borrowed Arabic words or names that should remain unchanged. The Pashto lexicon, on the other hand, uses the Arabic kaf to represent /k/ about 20% of the time. An extended Pashto normalization subroutine is being developed for the lexicon.

Pashto Lexicon:

Arabic digits 0660-0669 > Farsi digits 06F0-06F9
Farsi yeh 06CC ی > Arabic yeh 064A ي
Farsi /g/ 06AF گ > Pashto /g/ 06AB ګ
Arabic /k/ 0643 ك > Farsi /k/ 06A9 ک

Currently, the same spelling normalization is applied to both Dari and Pashto. However, the distribution of yeh variants in Dari might argue for preserving both variants. In the Dari data, the yeh variants occur with similar frequency, mostly distributed in sections, so it appears that different writers prefer one or the other consistently, as opposed to one character being an error.

Dari exhibits an alternation between Farsi /k/ 06A9 ک and Arabic /k/ 0643 ك . As in Pashto, however, the Arabic form is so rare that it probably represents borrowed words, which should remain unchanged.

French Gigaword: A separate spelling normalization scheme was developed for the French gigaword data. The program, giganorm.pl, removes control characters and converts ligatures to character sequences. For example, the ligature FB02 ﬂ becomes the two-character sequence, fl. Subsequently, the program removeForeignWords.pl is applied to remove sections of non-Latin characters. This program considers context, in order to retain non-Latin characters in names like *Jože*. A problem remains with the removal of symbols in formulas, such as *α=0.05*.

The removeForeignWords.pl program was extended to allow the user to specify which characters to retain, based on the character ranges for the following set of languages: English, French, Arabic, Russian, Greek, Hindi, and Chinese.

Other unusual text was identified in the gigaword data, including small caps encoded in the Unicode private use range and ciphered text that is apparently created during PDF file extraction. Ciphered text displays properly in the PDF file, but when extracted shows a pattern of letter substitutions, e.g., "The problem created by the current legislation" becomes "7KH SUREOHP FUHDWHG E\ WKH FXUUHQW". At least two distinct mappings were found in the data, and some lines contained a combination of ciphered and plain text. Given the minor amount of text affected, it was decided not to normalize these sections.

Unicode: A review was made of the 2010 changes to the Unicode code charts, version 6.0.0, to check for any needed changes in the normalization programs.[6]

Tokenization: Tokenization separates words and punctuation in order to better capture patterns of word use. This section reports on the tokenization programs developed for Urdu, Pashto, and French.

General Tokenization: The LDC tokenizer for English and French was improved in several ways; this created a general tokenization process which was then adapted for other languages. Six steps were initially considered:

0: Remove or replace repeated punctuation marks

1: Tag email and Uniform Resource Locator (URL) addresses to prevent changes

2: Replace HyperText Markup Language (HTML)with Unicode (e.g., &apos; becomes ')

3: Metric conversion

4: Digit-to-word conversion

5: Tokenize (put spaces between words and punctuation)

Of these, steps 0, 2, and 5 were implemented in the tokenizer. Step 1, protecting email and URL addresses, was split off into a separate program, tagurl. Step 4, digit-to-word conversion, was handled by writing separate programs for each language (see Section 0). Step 3, metric conversion, was discarded as having too many complications (e.g., some translations from English to French kept the digits the same, treating *1 mile* and *1 kilometer* as metaphorical equivalents, while other translations made precise calculations to convert them.)

The order of steps was eventually revised. The "025" tokenization was applied to the English and French TED talk data; a "205" tokenization was used for the English and French WMT11 data.

Lowercasing generally occurs after tokenization. A problem was identified in the application of lowercasing to acronyms. A revised lowercasing program was written that retains uppercase in words that have two or more capital letters, such as CNN, or in hyphenated words like Coca-Cola.

Section 0 discuss specific adaptations that were made for the tokenization of different languages.

[6] http://www.unicode.org/versions/Unicode6.0.0/

<u>*Urdu Tokenization*</u>*:* Hyphens were sometimes found in place of the Urdu full stop character 06D4 ۔ in both the Basic Travel Expression Corpus (BTEC) and IR data sets. Hyphens at the end of a segment were converted to stops, while hyphens in the middle of a segment were compared to the English translation to see if there was a sentence break or meaningful hyphens in words like ای-میل/ay-myl/ "e-mail". A test was made to see how the full stop characters were handled by the Systran translation system. It was discovered that an additional step of replacing all full stops with Latin periods allowed Systran to generate better translations.

<u>*Pashto Tokenization*</u>*:* The Pashto tokenizer supplied by the LDC lacks some of the punctuation characters found in the Sada-e-Azadi Pashto data, including ؟ 061f and ، 060c. A new Pashto tokenizer was created based on the SCREAM lab English tokenizer, applying character classes to include all punctuation marks as specified in the Unicode charts. The tokenizer was also adjusted to handle some word-initial attached punctuation that was found in the Pashto data. Applying this new augustTokenizerPlusLeadingDot.pl program showed a one-point Bilingual Evaluation Understudy (BLEU) score improvement vs. untokenized Pashto data.

<u>*French Tokenization*</u>*:* This section describes improvements that were made to the LDC French tokenizer and detokenizer, including better handling of punctuation and separate treatment of meta-data lines.

General French Tokenization: Improvements were made to the LDC French tokenizer, allowing better handling of contractions, abbreviations, apostrophes, and quotation marks, including right- and left- quotation marks as well as the French guillemet marks, « ».

An exception list was made for French words that include the apostrophe but should be written as one word (e.g., *aujourd'hui*). Improvements were made to the list of French abbreviations that allow a period without sentence-final spacing, such as M. for *Monsieur* or Prof. for *Professeur*.

Letter-number combinations were examined for possible tokenization. Some letter-number combinations are errors, like *24ans* '24 years' and *8Avril* 'April 8th', but other French combinations are correctly written together, as in the following abbreviations:

French Abbreviation	French Phrase	English Translation
1er	premier	1st
2e	deuxième	2nd
2h30	2 heures 30 minutes	2 hours and 30 minutes

Therefore, it would be a mistake to separate all letter-number combinations.

French TED Talk Tokenization: The tokenization process was revised to take better advantage of the TED talk meta-data, removing non-informative lines (talk id, keywords), and removing tags and tokenizing informative lines (titles, descriptions). Since the descriptions comprise multiple sentences, these were also separated these using an existing SCREAM lab program that splits parallel text when each side contains the same number of sentence punctuation marks.

A problem was identified in the TED talk data, in which HTML tags were sometimes attached to the text, instead of occupying separate lines, as in: applause</transcript>.

French Detokenizer: The LDC detokenizer.perl program failed to remove spaces around apostrophes involving the last two words of a French sentence, due to an error specifying the last word to examine. In English, the program examines all the words, since the apostrophe may attach to the final word, as in *can 't*. In French, after tokenization, the apostrophe can only occur on the penultimate word, as in *c' est*. The detokenizer, however, was only considering the third-to-last word for French. This error was corrected in a revised program, detokenizer3.perl.

Other problems were identified that are difficult to handle in detokenization. These include multiple-line quotations and detached sentence punctuation. The original program re-attaches quotation marks to words by looking for pairs: the first mark attaches to its following word, and the second mark attaches to its preceding word.

input	detokenizer.perl output
abc " quotation is here " def	abc "quotation is here" def

However, in the current data, there are longer quotations that are split across lines, creating unpaired quotation marks:

abc " a long quotation begins here and
ends here " def

A single quotation mark in the middle of the sentence cannot be accurately placed, but a sentence-final quotation mark can be assumed to belong to the preceding word. The revised detokenizer therefore left-attaches any sentence-final quotation marks.

Traditional French usage requires a space before sentence punctuation marks other than the period. The French text contains a mix of usage, in which some of these punctuation marks are attached to the last word of the sentence, as in English, and some are preceded by a space. The detokenizer works on the assumption that sentence punctuation is originally attached, with a space inserted by the tokenizer, which the detokenizer then removes. Given the variety in the French data, however, the detokenizer cannot consistently recover the spacing after tokenization.

Analysis of French Tokenization Output: A qualitative analysis of some of the MT output with and without SCREAM lab tokenization of the TED talk data shows that the tokenized data produced a better treatment of negation. The non-tokenized data tended to drop the word *pas* in the French negative phrase, *ne* [verb] *pas* "don't [verb]."

A comparison was made between the SCREAM lab tokenizer and the Massachusetts Institute of Technology (MIT) tokenizer. The programs differ primarily in the treatment of contractions and hyphenation, with the MIT tokenizer removing hyphens altogether. There is a trade-off here: Hyphen removal may alleviate data sparseness by conflating single words with words from hyphenated phrases; but hyphen removal also prevents the MT system from generating French constructs like *est-ce que* "is there?" or the grammatically-required hyphens in inverted phrases like *avez-vous* "have-you?" A variant of the MIT tokenizer was created that preserves hyphenation; tests showed that this change did not improve the MT.

2.2.3. Morphological Processing

Morphological variation can prevent the MT system from recognizing different instances of the same word. This section discusses ways to reduce morphological variation in Dari, Pashto, and French.

Pashto Morphological Normalization: Pashto morphological analysis is complicated by spacing variation and by the placement of adpositions. Pashto writers frequently run words together, particularly if the characters are non-joiners. For example:

Pashto	English (Literal)	English Translation
څلورکاله	four-years	four years
دڅلورو ورځو	in-four days	in four days
دافغانستان	of-Afghanistan	Afghanistan's

One normalization that might be profitable would be to split off the Pashto preposition د /d/ "of" when it is attached to a following word. A program could be written to detach the character د from unknown words, when the remaining word matches a dictionary entry. Another approach to resolving Pashto spacing variation is described in Section 0

The placement of adpositions also complicates Pashto morphological analysis. Pashto has clitic forms that occur before or after the stem, including several "ambipositions" that have both a pre- and post-positional component. For example, the sequence /ph/ __ /ky/ means “in ___”.

MT might be improved by splitting such morphemes from their stems by a rule-based program. The Humayoun Urdu morphological analyzer[7] was considered as a possible base for building a Pashto program. The Humayoun program is written as a modular component for use with the Functional Morphology toolkit, which is written in Haskell [10]. A review of the Urdu rules in Humayoun was conducted, along with a review of proposed Pashto rules as described in academic papers. In particular, Zuhra and Khan 2009 [11] delimit the morphological rules needed to implement Pashto noun phrase morphology within the Functional Morphology toolkit.

Another alternative was considered: the Grammatical Framework (GF) system[8], written in a language similar to Haskell. GF uses syntactic rules to parse multiple languages into a semantic interlingua. GF is easy to use because it is written according to linguistic principles, but it is also brittle: A sentence with an unknown word receives no parse output. There are references to newer versions which allow statistical adaptation, including the Multilingual Online Translation (Molto) system.[9] Another advantage of Molto is that there are existing Molto language resources such as grammars for Hindi and Urdu, and an extensive English wordnet-based lexicon. Finally, other tools were reviewed, including Apertium and the Natural Language Toolkit (NLTK).

Dari Morphological Normalization: While Dari does not exhibit as much morphological variation as Pashto, the Dari dataset does includes frequent instances of run-together words. For example:

Dari	English (literal)	English Translation
یک و نیم	one and half	one and a half
یک ونیم	one and-half	one and a half
دونیم	two-half	two and a half

[7] http://www.lama.univ-savoie.fr/~humayoun/UrduMorph/
[8] http://www.grammaticalframework.org
[9] http://www.molto-project.eu/

Dari grammars also indicate the possibility of variation in the pronominal system. For example, the combination /mn ra/ = /mn/ "I" + /ra/ (accusative) can be contracted to /mra/. The Dari dataset was reviewed and the presence of these pronominal contractions was confirmed.

French Morphological Processing: Output was reviewed for experiments using stemming and truncation programs in English to French translation. The output was reasonable but the stemming program appears to perform differently from its documentation: Some French inflectional endings such as -er and -ez were not stemmed, and some instances of final -s were incorrectly removed (for example, the conjunction mais "but" became mai).

2.2.4. Sentence Alignment

When parallel text is used to train MT models, a sentence alignment step attempts to relate each sentence in the first language with a corresponding sentence or sentences in the second language. This section discusses the detection and correction of errors in sentence alignment in French and Pashto.

Reviewed Sentence Alignment of Pashto and French Data: A word-for-word translation utility was used to confirm suspected sentence-alignment errors in Pashto/French data. The utility program was improved by adding Pashto characters for transliteration of unknown words, which helps in the detection of borrowed words.

Corrected Sentence Alignment in French Gigaword: The gigaword dataset exhibits two characteristics that require sentence splitting, involving the newline control character 2028 and the bullet character, • 2022 (and variants). The newline control character 2028 is used in the gigaword data to format the display of data such as addresses, which appear as multiple lines for street, city, and state. The 2028 character creates the appearance of new lines but the entire address is segmented as one line for MT. A controlSplit program was written that removes the 2028 character and splits the line at that point.

The gigaword data also contain bulleted lists, written on a single line with sections marked by various bullet characters such as • 2022. The SCREAM lab program, split_by_punct.pl, was adapted to create a split_by_bullets.pl program, which splits lines at bullet symbols, provided the same number of symbols are found in the French and the English versions of the line. The initial bullet symbol list includes the Unicode character range 25A0-27A4 and the symbol 2022. A second program, unbulletAll.perl, was written to remove the bullet symbols after splitting.

Corrected Sentence Alignment in French TED Talk Data: Several talks in the French TED talk data have sentence extraction errors. Apparently, when an English sentence is split across two or more lines and the corresponding French sentence is written on a single line, the sentence extraction algorithm then repeats this whole sentence to match the two English lines. Additionally, the French data contain some sentences with internally repeated phrases.

A program was written to identify the duplicate sentences, and to identify and automatically remove sentence-internal phrases. The program was applied to the 50 sentences with the worst word alignment scores, and 16 talks were identified that required substantial correction. The phrase removal program was refined to protect repeated French phrases if they are matched by repeated phrases in the English sentence. A parameter was added for the length of the repeated phrase, with an offset to account for the overall shorter length of English phrases. Hand editing was required to complete the restoration of these talks.

2.2.5. Word Alignment

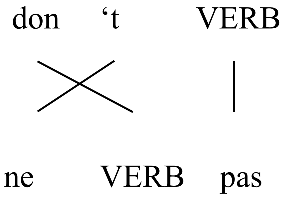
During the MT process, an attempt is made to assign links between words that are translation pairs in a given sentence. Errors at this stage lead to incorrect phrase table entries, which can then generate poor translations of future data. This section discusses the analysis of word alignment errors in the French system, and also reports on the use of human translators to edit word alignments in French and Urdu.

French Data: The French data exhibit problems in word alignment when the translation does not match the original English, and in negation constructions with the discontinuous phrase *ne __ pas*. Repeated words were investigated but ruled out as a source of word alignment errors.

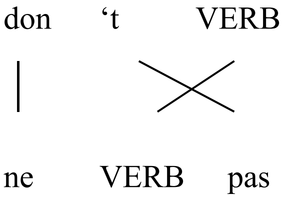
Translation Errors: Word alignment scores were considered for the TED talk data, and 66 of the worst-aligned sentences were examined to determine the type of error involved. Sentence alignment errors accounted for errors in 16 sentences (see Section 0). For the remaining sentences, the poor word alignment derived from factors such as English music lyrics repeated verbatim in the French translation and paraphrasing in the French translation. Paraphrasing often occurs if the English sentence contains fragments or false starts that are then cleaned up by the French translator. Lengthy sentences with an accumulation of French function words that align to NULL also scored poorly, but do not represent alignment errors.

Negation: Problems were also observed with the word alignment of negative phrases when negation involves the discontinuous phrase, *ne __ pas* on the French side, and a contraction with auxiliary verb, *don'‘t* , on the English side. After tokenization, we have *ne [verb] pas* and *don 't [verb]*; since the English auxiliary is not meaning-bearing, there is no French element to align with *don*. The extraneous element *don* can end up linked to the French verb or to the negative particle *ne*. The preferred alignment links the English negation element *'t* to both *ne* and *pas*.

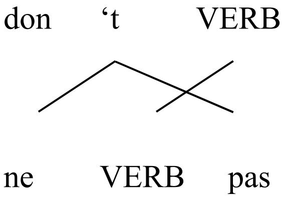
Incorrect



Incorrect



Correct



Two ways to improve these alignments were considered: text-editing to remove the auxiliary *don* before alignment takes place, and post-alignment editing to de-link English *don* from French *ne* and enforce a link between French *ne* and English *‘t*. Programs were written to test each method; results were mixed. In particular, the text-editing version sometimes caused the *ne* element to drop out, which is not desirable.

Repeated Words: A number of repeated words were noticed in the gigaword data. A program was written to examine these words to see if they might derive from hesitations in the original audio. Instead, many repeated words were found to be legitimate vocabulary items (***great great*** *grandmother,* ***win win, trillion trillion***) or legitimate grammatical constructions (*what we believe* ***in in*** *Canada*).

English-to-French Word Alignment Editing: A human translator was used to edit the automatically created English/French word alignments. The first 2000 sentences of the TED talk training data were examined by the translator and changes were made as needed. Guidelines were developed for the alignment of idioms, in which two phrases represent the same meaning across the languages, but the individual words do not correspond: These are best handled by aligning all component words in one phrase with all component words in the other phrase. Metrics were developed to evaluate the editing process, including counts of the number of links changed and the percentage of words left unaligned. Initial experiments were run using the hand-edited alignments.

The potential cost to align the entire IWSLT alignment file was calculated and found to be prohibitive. Initial metrics were developed to decide which sentences would be most valuable to send to a human editor for alignment editing. The existing human-edited alignment files were examined for variables such as the length of the sentence, the percentage of links changed, the automatically-assigned alignment score, and the percentage of links that match an existing dictionary entry (calculated from an existing SCREAM lab program).

Alignment Exploration: Alternative methods of word alignment were reviewed, including the use of anchors, in which single-word sentences or borrowed words are aligned, and then surrounding words are aligned iteratively. An initial word alignment utility was written that detects sentence pairs with one unaligned word in each language and then revises the alignment file to link those words.

A variation of the existing SCREAM lab lexicon matching program was written in which sentences which contain a high percentage of lexicon-validated links are considered to have good alignments, and the remaining links in those sentences are used to generate additional lexical entries. This process can then be iterated, drawing in more sentences that are lexically-validated. This program can be varied by adjusting the threshold for the percentage of matching links and by adding filters to prevent some types of links. An interface program allows a human editor to review the new entries.

An examination of the new word lists showed some improbable alignments; some of these were due to errors in the hand-editing, and others derived from problems in the automatic alignment process, particularly in the alignment of phrases. For example, given a sentence pair with the English phrase *a baker* corresponding to the French translation *boulanger*, the automatic alignment process might link only the article instead of the whole phrase: *a* -- *boulanger*.

2.2.6. Phrase Table Problems

An analysis of the French/English phrase tables from the TED talk data shows that there are English words among the French. This can occur when the parallel text contains names or direct quotations that are unchanged in translation. For example, the name *Gladys Knight and the Pips* is recorded unchanged in the French parallel text, creating in the phrase table the possibility that English *and* translates to *and* in French. This then allows the MT to use *and* as a conjunction within a French sentence, where the French word *et* should have been used instead.

2.2.7. Word-Level Normalization

The presence of alternate word forms contributes to data sparsity. This section reviews the normalization of alternate word forms through digit-to-word conversion in Dari, Pashto, English, and French, as well as the generation of alternate French tense forms.

Digit-to-Word Conversion: Digit-to-word conversion is a kind of normalization that helps MT in two ways. First, by converting digits to the corresponding numeral words (e.g., *2* becomes *two*) we reduce the amount of variation in the data. Second, digit conversion improves word alignment: Without number conversion, a single numeral word may become aligned with an entire digit sequence e.g., (three|1397).

Digit-to-word conversion goes beyond simple mapping (*2 > two*), since the same digit may be read in different ways in different contexts. In English, for example, the sequence *123* might be read as "one hundred and twenty three" by itself, but as "one two three" when part of a decimal *0.123* or a phone number, *555-0123*. This conversion has to take into account the specific patterns in each language.

Existing number conversion methods were reviewed, including the icu4j Java module for internationalization. It was determined that it would be useful to create independent SCREAM lab conversion programs for various languages. The SCREAM lab program for English number conversion was based on the CPAN Perl module Lingua::EN::Numbers::num2en, and extended to handle the examples found in the English training data. The SCREAM lab program for Urdu number conversion was based on the CPAN Perl module for Chinese, Lingua::ZH. Current efforts extend these programs.

Dari: Lists were found for the numeral words for 1-100 in Farsi; these can be used for Dari, since Dari and Farsi are closely related. Existing SCREAM lab number conversion programs for English and Urdu were consulted in the construction of the Dari/Farsi version, with certain modifications. English and Urdu make up the hundreds from pieces (e.g., *200 = two hundred*), but Farsi has distinct words for each of the hundreds (*200 = devist*). Dari uses some combined words and some distinct words for the hundreds.

For each number that is expressed in digits, the program counts up the number of tens, hundreds, thousands, millions, and billions. At each level, that amount is converted to words using a digit-to-word mapping for the values 1-99, with the addition of the appropriate Dari unit words and conjunctions. A special mapping is applied for the hundreds words.

Digits	Dari Words	Literal Translation
254	دو صد و پنجاه و چهار	two hundred and fifty and four
454	چهارصد و پنجاه و چهار	four-hundred and fifty and four
3254	سه هزار دو صد و پنجاه و چهار	three thousand two hundred and fifty and four

Specific handling is added for decimals, currency expressions, and phone numbers.

Digits	Dari Words	*Literal Translation*
$5.20	پنج دالر بیست سینٹ	five dollar twenty cent
555-8427	پنج پنج پنج هشت چهار دو هفت	five five five eight four two seven
3.25	سه اعشاریه دو پنج	three point two five

Further research is needed on Dari to see if phone numbers or decimals should be read as single digits, as shown here, or as groupings of larger numbers (as is the case in Urdu, for example).

For details on decimal conversion, see the discussion in the English section 0 below.

Pashto: Lists were found for the numeral words for 1-100 in Pashto; these standard spellings were compared to the actual spellings found in the training data. Pashto numeral words have inflected forms for different grammatical functions such as subject vs. object. Pashto also has highly variable spelling in general, so there are many different forms of the numeral words in the training data. The unit words also show variation.

Some of the variants reflect case endings (the oblique case may surface as *-w* or *-h*), while others show spelling variations in the stem forms. The training dataset exhibits the following variations:

Digit	Pashto	Transliteration	Notes
2	دو	dw	
	دوو	dww	with ending -w
	دوه	dwh	with ending -h
11	یوولس	ywwls	stem 1
	یوولسو	ywwlsw	stem 1 with ending -w
	یولس	ywls	stem 2
	یولسو	ywlsw	stem 2 with ending -w

English	Pashto	Transliteration	Notes
million	ملیون	mlywn	
	ملیونه	mlywnh	with ending -h
	ملیونو	mlywnw	with ending -w
	میلونه	mylwnh	stem change, with ending -h

The Pashto number conversion program follows the design laid out for Dari, above (Section 0), with special handling of variation. The Pashto dataset was used to create a frequency list for the number word variants; the number conversion program randomly selects one of these variants, at a rate designed to match the frequency in the training data. A revised version of this program is planned that would create a lattice of possible forms for each number, along with a weighted probability taken from the frequency of each form in our training data.

For details on the treatment decimal conversion, see the discussion in the English section 0 below.

English: The style of English used in the English/Dari/Pashto data (from the Sada-e-Azadi news site) differs from previous English training data, in that the numbers contain a mix of American and European style punctuation. Punctuated numerals are therefore ambiguous.

Puncutation	American	European
1.234	decimal	thousands
1,234	thousands	decimal

The original SCREAM lab English number conversion program interprets digit punctuation in the American style. For the Sada-e-Azadi data, this program was adapted to make informed guesses as to the interpretation of digit punctuation, based on the number of trailing digits, as shown below. This procedure is also followed in the Dari and Pashto number conversion programs.

Punctuation	Best Guess
x,xxx	thousands (unless it has a leading 0)
x.xxx	thousands (unless it has a leading 0)
x,xx	decimal
x.xx	decimal

Another variant of the English number conversion program was made for the gigaword dataset, writenum18_Gigaword.pl. This version prefers the large number reading to the European phone number reading for numerals punctuated with spaces (e.g., *38 000 > thirty eight thousand*, not *38 000 > three eight zero zero zero*). Finally, an alternate version was created which allows the phone number reading in just American-formatted phone numbers.

Number Conversion in the French TED Talk Data: The French number conversion program was based on the existing SCREAM lab English number conversion program. Changes were made to handle the French punctuation of numerals, in which a comma indicates a decimal, and thousands are marked with a space or a period.

The French data supplied on the TED talk website was created by human translators working from the original English files. These translators typically converted English units to metric units. This will create problems for word alignment when the system attempts to align, for example, *20 miles* with *32 km*, creating an entry for 20|32. An attempt was made to prevent these mismatches by converting the English training data to metric units as well. The program identified numeral-unit phrases and then applied the CPAN Perl modules Math::Units::convert and Lingua::EN::Numbers::words2nums.

However, there are difficulties in applying metric conversion. Human translators may use approximations, depending on the context; sometimes *un kilometer* is used as a metaphorical equivalent to *a mile*, for example. Also, some English units are ambiguous: *pounds* might need to be converted to *kg*, or it might refer to currency. Given these problems, and the relative infrequency of the metric conversion items, it was decided not to include metric conversion for the English/French data.

A distinction was made between ordinary decimals (which are read as large numbers in French) and decimals with leading zeroes (which are read as digits). Certain abbreviations were added, converting the abbreviation *h* to *heures* 'hours' in time expressions and converting the degree sign to the word *degree*. Finally, code was written to generate French fraction words.

Input	**Output**	**Literal Meaning**
3,125	trois virgule cent vingt cinq	three point one hundred twenty five
3,025	trois virgule zero deux cinq	three point zero two five
19h50	dix neuf heures cinquante	nineteen hours fify (minutes)
1er	premier	first
2e	deuxième	second
1/2	une moitié	one half
1/3	un tiers	one third

The French number conversion program anticipates European style decimal punctuation, but both European and American style decimal punctuation are present in the French data. Future work may need to apply a contextually-based interpretation of decimal punctuation, as described in Section 0

European	**American**
80 000	80,000
2,5	2.5
24 000 US$	$24,000
1.200,17 US$	$1,200.17

A review was made of the results of translating English to French with number processing, in which the digits are converted to words for both the English and the French data. In this process, for example, the English digits *20* become *twenty*, while the French digits *20* become *vingt*, and the system then learns the association between the words *twenty* and *vingt*. The existence of borrowed English number words in the French caused problems. For example, the presence of the named entity, *Nine Inch Nails* in both the English and French led the system to create a phrase table entry relating *nine* in English to *nine* in French. (See Section 0)

Generation of Morphological Variants: Another approach to data sparsity with inflected forms is to generate the missing forms. This section reports on the automatic generation of alternate tense forms for French verbs, which appear in one past tense form in the training data but another past tense form in the test data.

French Tense Conversion: It was observed that the Out Of Vocabulary (OOV) words in the French data included a large number of verbs in the simple past tense, which is typically found only in formal written language. Since the training data is from the spoken europarl text, it contains fewer of these forms. A tense conversion program was developed to change the compound past tense forms to simple past, in hopes that the enriched training data would then provide a basis for translating these forms.

The tense conversion algorithm gathers information from the compound past form and creates the simple past form. For example, the compound past form *j' ai parlé = I have spoken* becomes, in simple past, *je parlai = I spoke*. The program determines the person and number from the auxiliary verb, and the conjugation from the characteristic vowel of the main verb. Person, number, and conjugation together determine the form of the suffix in the simple past form.

These factors were addressed: choice of auxiliary verb (*be* or *have*), irregular verb conjugations, intervening adverbs, and changes in the contraction of pronouns. For example, the pronoun *je* = *I* contracts to *j'* before the vowel in *j' ai parlé = I have spoken*, but in the simple past, with no auxiliary, the non-contracted form, *je parlai = I spoke* is required.

The application of the tense conversion program improved the translation qualitatively (simple past forms are now recognized), but failed to increase the BLEU score.

2.28 Separating Data into Domains Using Metadata

MT is sometimes improved by creating domain-specific models. This section reports efforts to separate data into domains using metadata such as the TEDTalk keywords.

Keyword Topic Sorting for French TED Talk Data: The possibility of sorting the TED talks by keyword topics was considered. The TED website assigns keywords to the talks at several levels. Initially, the talks were supposed to be about Technology, Entertainment, or Design. To these keywords were added Business, Science, Culture, Arts, and Global Issues; smaller divisions are also provided. A single talk can be tagged with multiple keywords. An examination of the distribution of the main keywords among the talks in the training and dev data showed that, even at the T-E-D level, there are significant overlaps. A program was then written to automatically sort talks by the Technology keyword, creating a high-level division.

Readability and Interagency Language Rating (ILR) Scale: A review was made of literature on readability, possible correlations between readability metrics and the ILR scale, and the question of whether BLEU scores may be used to help detect ILR level.

2.2.9 Machine Translation Output

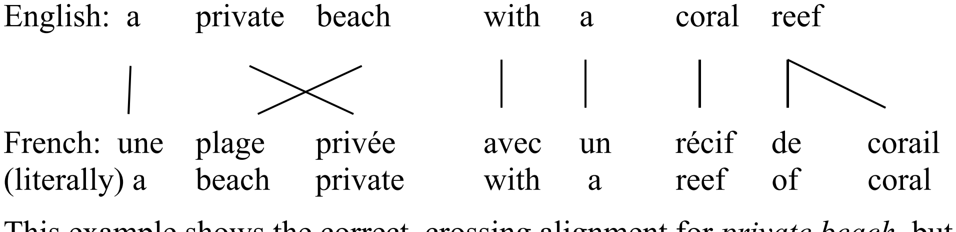
This section reports the results of examining the French MT output for vocabulary gaps and syntactic errors.

English Out of Vocabulary Words: The MT system lets unknown words pass through unchanged. Examining these untranslated OOV words can identify problems in the system. It is difficult to detect OOV words when translating from French to English, since the languages have similar alphabets. A method was developed to help identify English OOV words. The first step identifies words which are unchanged in the course of the translation; the second step applies the spelling correction program Aspell to distinguish leftover English words from words that happen to be the same in French and English (such as *construction*).

Some untranslated words cannot be identified by this method due to accidental homography between some French and English words (for example, English *as* and French *as* "have" 2nd person singular).

Some words were not spelled correctly in English or French. A few of these were due to the omission or misspelling of French accented characters. Other OOV words involved hyphenation, particularly words which contained grammatically-induced hyphenation, such as *devrions-nous ?* "must we ?" Here, the MT system can recognize each word independently, but has not seen the hyphenated form that is generated when the sentence is placed in the interrogative.

Word Alignment Errors with French Noun-Adjective Constructions: The output of the initial English/French MT failed to reorder nouns and adjectives. French typically reverses the order of nouns and adjectives when compared to English, so *creative spirit* should become *esprit créatif* – but the system generates "*créatif l'esprit*", maintaining the English order. This was traced this back to a problem in the word alignment: the GIZA++ word alignment program frequently aligns the French and English words in the order they appear, instead of with the crossing alignment necessary for adjective-noun phrases. For example, GIZA++ makes the following alignment:



English:	a	private	beach	with	a	coral	reef	
French:	une	plage	privée	avec	un	récif	de	corail
(literally)	a	beach	private	with	a	reef	of	coral

This example shows the correct, crossing alignment for *private beach*, but an incorrect alignment for *coral reef*. This kind of error could be corrected with a lexical consistency check, by applying an existing SCREAM lab program that looks for lexical entries corresponding to the word alignment links.

2.2.10. Paraphrasing

MT scoring penalizes translations which contain the correct meaning but do not match the exact words of the reference sentence. Paraphrasing research addresses ways to automatically create variant sentences with similar meanings, for possible use in training or scoring of MT.

Paraphrasing with the Callison-Burch Software: Previous work on paraphrasing had created a program to interface with the Callison-Burch (CB) paraphrasing software [12]. Extensions were made to specify the number of changes allowed per sentence, and to prevent changes of certain types.

Previous work had implemented restrictions that prevent paraphrases that change punctuation or change a number word, as in the examples [.>*!*] and [*one* > *two*]. Other restrictions were added that prevent changes to digits [*2*>*1*], the paraphrasing of a word to punctuation [*and*>*!*], or the paraphrasing of a content word to a function word (article, conjunction, or preposition). This last restriction is necessary when using the LM to weight paraphrase choices, since the function words occur with higher frequency and are therefore preferred over more meaningful paraphrases. This restriction prevents paraphrases like [*of thinking* > *of*] and allows paraphrases like [*of thinking* > *of thought*].

Paraphrases were created for the WMT English reference translations, and the range of variation was examined using BLEU scores. Future work may use such paraphrased references to provide better scoring of translation data.

Paraphrased data were created for some of the IWSLT experiments, including paraphrased English data for the Chinese-to-English experiments, and paraphrased French data for the English-to-French experiments.

Paraphrasing with Lattices: A program was written to generate lattice representations of paraphrases for use in the Moses MT system. The lattice format allows the system to maintain alternatives for consideration throughout the translation process. For the lattice format, a sentence is represented with each word being an arc that links two nodes in a sequence; alternative wordings are indicated by additional arcs between nodes. When a single word alternates with a phrase, the lattice must have additional nodes to anchor the intermediate steps in the phrase.

For example, Figure 2 shows the addition of a paraphrase in which the acronym *MT* alternates with the full phrase, *machine translation*:

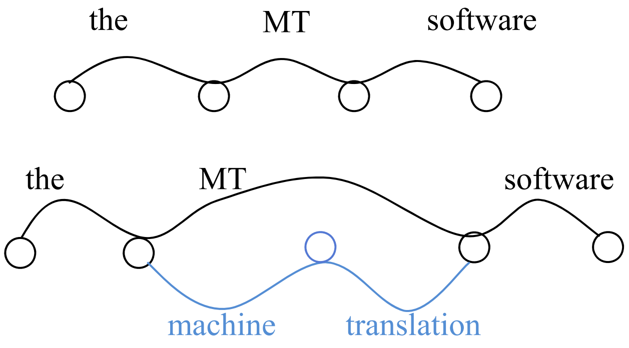


Figure 2: Lattices Showing the Addition of a Paraphrase

The CB paraphrasing software was applied to generate the paraphrases, which were then stored in a format that indicates the location of the words to be replaced. A second program generates any new nodes that are needed to represent replacement phrases, and interpolates these nodes into the original sentence. The edges of the lattice are assigned weights that indicate the likelihood of choosing each word. The paraphrase probability and LM probability values from the CB software were used to assign edge weights.

In addition to lattices, experiments were conducted using confusion nets. Confusion nets are further constrained so that each path traverses every node. A program was written to convert the paraphrase lattices from the Moses Python Lattice Format (PLF) to the HTK Standard Lattice Format (SLF). This enables the use of the SRILM tool to reduce the lattices to confusion nets, a form which provides more flexibility for optimization.

The derivation of confusion nets from the Moses output search graph and/or the Moses output word graph was also considered. These structures record the alternatives that were considered during the MT process, and could be used as input for further processing. It was determined that this format does not provide all the information required for the generation of the confusion net.

Lattice variations were created including single paraphrases vs. sets of nbest paraphrases, paraphrases with and without the guidance of a LM, paraphrases with heavier weights given to the original words, and unconstrained vs. syntactically-constrained paraphrases. The syntactically-constrained option is provided by the CB software, and refers to paraphrases that are derived from sentences in which the two phrases share the same syntactic label (e.g., both are noun phrases or both are verb phrases). Experiments were run with various options, and indicated that the best combination was created using syntactically-constrained paraphrases, without the LM, in the three-best condition, with a higher weight assigned to the original phrases. However, even the best-scoring experiments with the paraphrases did not exceed the baseline scores.

Two approaches were used to attempt to improve the paraphrase scores: combining LM and paraphrase probability scores, and improving the use of the LM scores.

The lattice generation program was revised to retain both LM and paraphrase probability scores. Then the SRILM lattice-tool parameters acscale and lmscale were used to create separate confusion nets for each score. These confusion nets can be combined if the resulting confusion net has the same structure, but sometimes the structures differ. Future work will consider how to modify the behavior of the SRILM lattice-tool program to create combined confusion nets.

A second method was considered to maintain and optimize separate scores for paraphrase probability and LM probability. First, a confusion net is generated based on the paraphrase probability scores; then, the trigram LM scores are calculated for each alternative word in the confusion net. This method is less than ideal, since it still gives preference to the paraphrase probability structure.

A domain-specific LM was created and applied to guide the selection of paraphrases. Usually a domain-specific model will give more useful results. However, using a LM based on the TED Talk domain creates a mismatch with the CB paraphrase tables, which were derived from Europarl, and contain formal language and foreign words not found in the TED Talk data. Normally, such paraphrases are discounted by the LM, which assigns only a small backoff probability to unknown phrases. However, because paraphrasing involves low overall probabilities, this backoff probability is sometimes higher than the probability of the known phrases, causing the model to select the unknowns instead of discarding them.

A review was conducted of the backoff and interpolation algorithms used to discount LM probabilities, in hopes of adjusting the probability assigned to unknown words. As a short-term solution, a filter was created that identifies the LM score that would be assigned to a nonsense word, and prevents the selection of paraphrases that receive that score. This allows more meaningful paraphrases to surface as the highest-scoring items.

The unknown words were examined. Many of the unknown words are British spellings that would be known to the TED LM in their American versions. The Varcon spelling conversion program was applied to the paraphrases in order to generate the American spellings. For example, the paraphrase *work* > *labour* is filtered as an unknown, but after spelling conversion, the program derives the useful paraphrase *work* > *labor*. Tokenization was another source of unknown words, when the tokenization differs between the paraphrase generation system and the LM file.

Paraphrasing with the Parex Software: The Paraphrase Extractor (Parex) program[10] was considered [13, 12]. A comparison was made between the CB software, which uses suffix arrays, and the Parex software, which uses suffix tries. A test set of paraphrases was created, using French/English parallel text to derive both French paraphrases and English paraphrases. Parex uses the phrase table created by the Moses translation software to derive pivot paraphrases. For example, given the following hypothetical phrase table entries, the program would identify *cat* > *kitten* as a possible English paraphrase, based on the presence of the pivot word *chat* "cat" in both entries.

English	French
cat	chat
kitten	chat

It was noted that Parex might be useful in deriving domain-specific paraphrases, since it can take any phrase table as input.

Paraphrasing with Meteor/NEXT Software: The Meteor/NEXT paraphrasing software was considered [13]. A problem was identified in the Meteor/NEXT paraphrase tables, in which some paraphrases are assigned a probability greater than one. Correspondence with the author of the software confirmed that this is a bug in the program. Pending correction of the bug, a filtering program was written to filter out paraphrases with impossible probability scores and to filter out paraphrases that have unknown characters due to encoding problems.

Paraphrasing with the Joshua/Thrax Software: The paraphrasing capability of the Joshua/Thrax program[11] was considered [14]. Joshua is a hierarchical phrase-based translation system, within which the subsystem Thrax operates to extract grammar rules. Thrax is usually used to extract a translation grammar (e.g., English>French), but can also be set to extract paraphrases. Syntactically-constrained paraphrases can be created by applying a parser for the target language.

[10] https://github.com/mjdenkowski/parex
[11] http://joshua-decoder.org/

The Thrax paraphrase system was installed and tested; a Hadoop system was installed to support the paraphraser. Initial problems with Thrax were solved via correspondence with the software's authors.

Joshua was used to create baseline en>fr and fr>en translation systems. Next, Thrax was used to create English paraphrases, using Joshua's built-in Berkeley parser. Then, Thrax was used to create French paraphrases, using an external Berkeley parser trained for French.

Paraphrasing and Plagiarism: A review was made of literature on the use of paraphrasing and comparable text, and of literature on the related field of plagiarism detection. Metrics and corpora for plagiarism detection were identified for possible application to paraphrasing.

2.2.11. Semantic Language Model

The source code, as delivered, contained errors that resulted in random alterations of the model. The expected input file format of 'ngram-count' was line based and made assumptions which were incompatible with the standard SRILM utility of the same name. The program was modified to match SRILM's file format and the document boundary specification was enhanced to accept escaped character sequences, which provided backward compatibility without inhibiting future changes.

2.2.12. Continuous Space Language Model (CSLM)

The CSLM tools had stability issues stemming from multiple segmentation faults. Further testing uncovered previously hidden problems with the way CSLM was mapping its own vocabulary to the SRILM model's vocabulary, which may have been giving erroneous results. These issues were fixed and the tools were enhanced to accept sentences which are smaller than the neural networks input dimensions. Documentation was updated and basic error checking was added to the 'ngram-count' program as well.

2.2.13. SRI Language Model

The Maximum Entropy and Random Forest extensions, with their support libraries, were integrated into the standard build process for SRILM. Errors in the Maximum Entropy extension were found and fixed. Additionally, a generalized version of the SCREAM Laboratory's mix-lm extension patch was created, allowing for an arbitrary number of models to be specified.

2.2.14. MIT Language Model

The output format was corrected to match the standard SRILM output, with back-off weights always being printed, regardless of value. The standard command line parser was also modified to accept a '-' to be specified as an input source, which allows for standard input, and therefore piped input, to be read instead of a regular file.

It was found that the perplexity optimization routines, in trying to save memory, were reusing the probability vector as temporary storage space for discount values. When using a mask, these values were not always overwritten with valid probabilities, causing errors. The solution to this problem is to use a separate vector for the discount values, which increases memory usage but gives a correct result.

2.2.15. Recurrent Neural Network Language Model

The Recurrent Neural Network LM contains many custom implementations of standard library functions and Basic Linear Algebra Subroutine (BLAS) calls. Replacing these custom implementations with the standard functions resulted in a 14% increase in word throughput. Several issues were fixed which now allow larger data sets to be used, where previously they could fail with segmentation faults or data corruption.

Additionally, an effort was made to create a direct port of the ‘rnnlm’ tool to the General Purpose Graphics Processing Unit (GPGPU). Much of the code is serially dependent, which makes GPGPU processing less than optimal, and memory transfer overhead was significant. The overall performance was slightly worse than the regular version on the CPU. A complete redesign of the tool could yield better results, but the effort would take a non-trivial amount of time.

2.2.16. Chaski

Chaski was updated to work with an updated version of Hadoop. A portion of the main Chaski tool was rewritten to accommodate long running jobs. Previously, if the job finished too long after starting, its metadata would have already been archived, causing a metadata lookup to fail.

The tool suite’s performance was analyzed against various sized data sets. A ten million sentence dataset completed on six Hadoop computation nodes in two days, which is not great, but performance is expected to increase with the number of nodes in the cluster. When run against a corpus of approximately 110 million sentences, Hadoop failed due to unspecified resource exhaustion. It is unknown if adding more nodes to the cluster would alleviate this problem.

2.2.17. Moses

Memory leaks in the Syntactic Language Model portion of Moses were found and fixed. This was preventing the processing of input data of any appreciable size. The build system was also modified to allow for static compilation of Moses binaries by removing a dependency on libSegFault for static builds. This approach does not reduce runtime functionality.

2.2.18. LatticeMERT

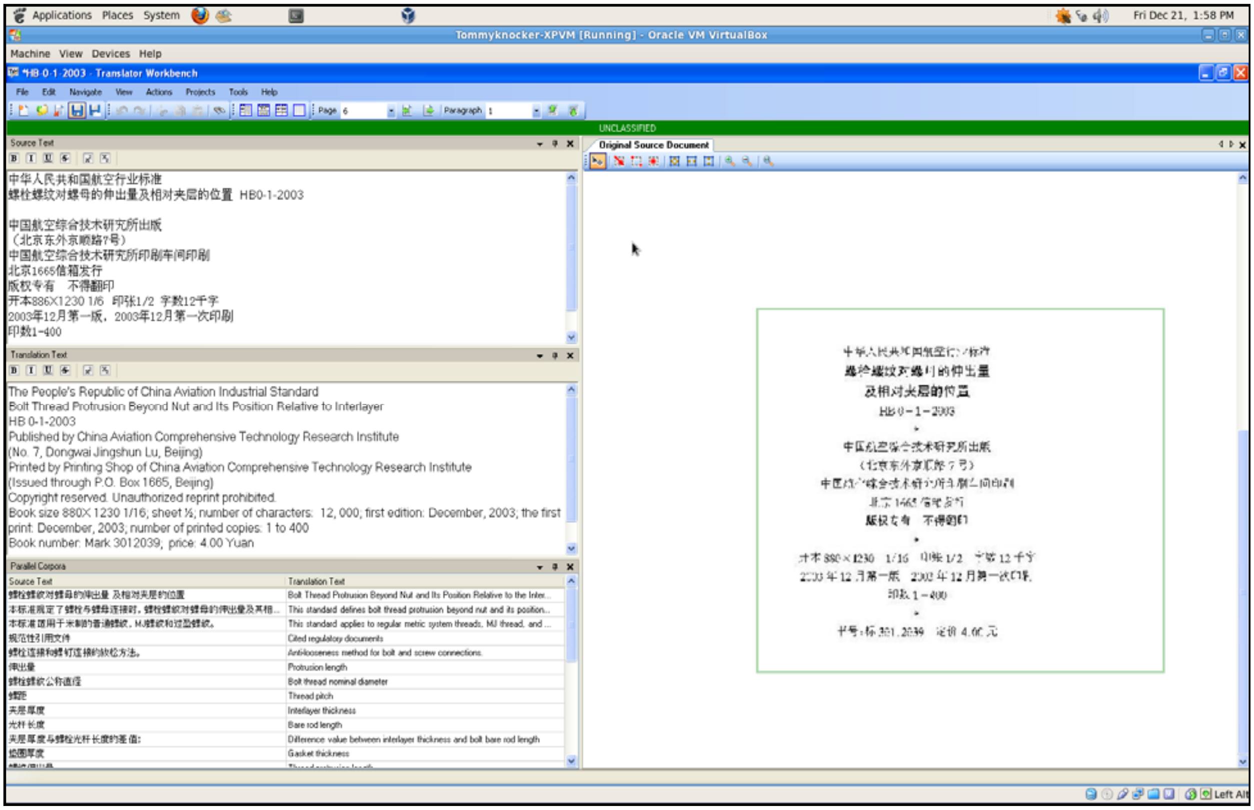
The format of Moses’ graph output changed slightly and LatticeMERT stopped being able to read the graphs. The graph reader was modified to use a more robust parsing routine when retrieving a phrase from the file, maintaining compatibility with both the old and new graph formats. It also now gives errors when parameters specified on the command line are known to be incorrect.

2.2.19. Translator Workbench Output

The Translator Workbench (TW) is a software tool developed by Northrop Grumman for assisting human translators in maintaining document layout and refining source and translation pairs of text for parallel corpora. The TW interface is shown in Figure 3.

The parallel corpora output by TW is formatted in Extensible Markup Language (XML) but for training LMs in the SCREAM Lab the text needs to be rendered into plain text.

A collection of scripts were developed in PHP to gather each pair of text files and scan through them line by line and extract the XML formatting and tags. The files were reconstructed into plain text with new line breaks and the line count of each pair of files was compared. If the line counts matched then the pairs of files were saved out to a new directory. If the line counts did not match after the formatting then they were left alone to be checked for formatting errors or discrepancies and, if possible through an easy fix, corrected and run back through the script for processing. The resulting plain formatted parallel text is shown in Figure 4.


Applications Places System
Fri Dec 21, 1:58 PM
Tommyknocker-XPVM [Running] - Oracle VM VirtualBox
Machine View Devices Help
*HB-0-1-2003 - Translator Workbench
File Edit Navigate View Actions Projects Tools Help
Page 6
Paragraph 1
UNCLASSIFIED
Source Text
Translation Text
Parallel Corpora
Original Source Document
Left Alt

中华人民共和国航空行业标准
螺栓螺纹对螺母的伸出量及相对夹层的位置 HB0-1-2003

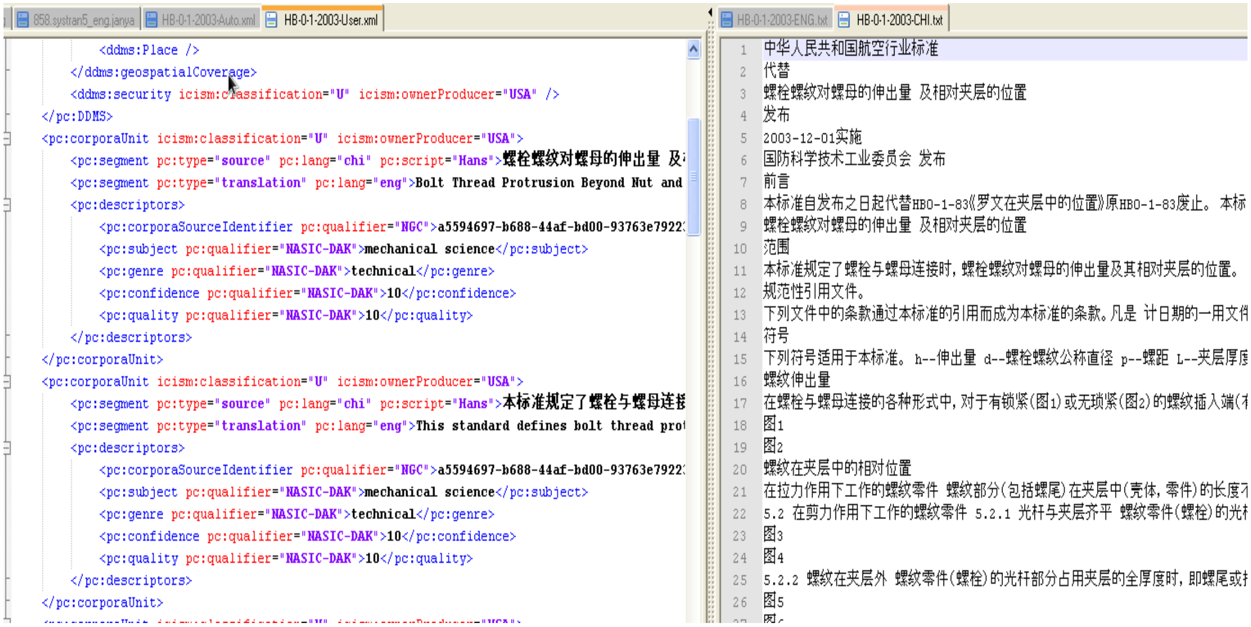
中国航空综合技术研究所出版
（北京东外京顺路7号）
中国航空综合技术研究所印刷车间印刷
北京1665信箱发行
版权专有 不得翻印
开本886X1230 1/6 印张1/2 字数12千字
2003年12月第一版，2003年12月第一次印刷
印数1-400

The People's Republic of China Aviation Industrial Standard
Bolt Thread Protrusion Beyond Nut and Its Position Relative to Interlayer
HB 0-1-2003
Published by China Aviation Comprehensive Technology Research Institute
(No. 7, Dongwai Jingshun Lu, Beijing)
Printed by Printing Shop of China Aviation Comprehensive Technology Research Institute
(Issued through P.O. Box 1665, Beijing)
Copyright reserved. Unauthorized reprint prohibited.
Book size 880X 1230 1/16; sheet ½; number of characters: 12, 000; first edition: December, 2003; the first print: December, 2003; number of printed copies: 1 to 400
Book number: Mark 3012039; price: 4.00 Yuan

Source Text	Translation Text
螺栓螺纹对螺母的伸出量 及相对夹层的位置	Bolt Thread Protrusion Beyond Nut and Its Position Relative to the Inter...
本标准规定了螺栓与螺母连接时，螺栓螺纹对螺母的伸出量及其相...	This standard defines bolt thread protrusion beyond nut and its position...
本标准适用于米制的普通螺纹，MJ螺纹和过盈螺纹。	This standard applies to regular metric system threads, MJ thread, and ...
规范性引用文件	Cited regulatory documents
螺栓连接和螺钉连接的放松方法。	Anti-looseness method for bolt and screw connections.
伸出量	Protrusion length
螺栓螺纹公称直径	Bolt thread nominal diameter
螺距	Thread pitch
夹层厚度	Interlayer thickness
光杆长度	Bare rod length
夹层厚度与螺栓光杆长度的差值:	Difference value between interlayer thickness and bolt bare rod length
垫圈厚度	Gasket thickness

中华人民共和国航空行业标准
螺栓螺纹对螺母的伸出量
及相对夹层的位置
HB 0－1－2003
中国航空综合技术研究所出版
（北京东外京顺路7号）
中国航空综合技术研究所印刷车间印刷
北京1665信箱发行
版权专有 不得翻印
开本 880×1230 1/16 印张1/2 字数12千字
2003年12月第一版 2003年12月第一次印刷
印数1－400
书号：标301.2039 定价 4.00元

Figure 3: An example of the Translator Workbench Interface



```
            <ddms:Place />
        </ddms:geospatialCoverage>
        <ddms:security icism:classification="U" icism:ownerProducer="USA" />
    </pc:DDMS>
    <pc:corporaUnit icism:classification="U" icism:ownerProducer="USA">
        <pc:segment pc:type="source" pc:lang="chi" pc:script="Hans">螺栓螺纹对螺母的伸出量 及
        <pc:segment pc:type="translation" pc:lang="eng">Bolt Thread Protrusion Beyond Nut and
        <pc:descriptors>
            <pc:corporaSourceIdentifier pc:qualifier="NGC">a5594697-b688-44af-bd00-93763e7922
            <pc:subject pc:qualifier="NASIC-DAK">mechanical science</pc:subject>
            <pc:genre pc:qualifier="NASIC-DAK">technical</pc:genre>
            <pc:confidence pc:qualifier="NASIC-DAK">10</pc:confidence>
            <pc:quality pc:qualifier="NASIC-DAK">10</pc:quality>
        </pc:descriptors>
    </pc:corporaUnit>
    <pc:corporaUnit icism:classification="U" icism:ownerProducer="USA">
        <pc:segment pc:type="source" pc:lang="chi" pc:script="Hans">本标准规定了螺栓与螺母连接
        <pc:segment pc:type="translation" pc:lang="eng">This standard defines bolt thread pro
        <pc:descriptors>
            <pc:corporaSourceIdentifier pc:qualifier="NGC">a5594697-b688-44af-bd00-93763e7922
            <pc:subject pc:qualifier="NASIC-DAK">mechanical science</pc:subject>
            <pc:genre pc:qualifier="NASIC-DAK">technical</pc:genre>
            <pc:confidence pc:qualifier="NASIC-DAK">10</pc:confidence>
            <pc:quality pc:qualifier="NASIC-DAK">10</pc:quality>
        </pc:descriptors>
    </pc:corporaUnit>
```

```
中华人民共和国航空行业标准
代替
螺栓螺纹对螺母的伸出量 及相对夹层的位置
发布
2003-12-01实施
国防科学技术工业委员会 发布
前言
本标准自发布之日起代替HBO-1-83《罗文在夹层中的位置》原HBO-1-83废止。 本标
螺栓螺纹对螺母的伸出量 及相对夹层的位置
范围
本标准规定了螺栓与螺母连接时，螺栓螺纹对螺母的伸出量及其相对夹层的位置。
规范性引用文件。
下列文件中的条款通过本标准的引用而成为本标准的条款。凡是 计日期的一用文件
符号
下列符号适用于本标准。 h--伸出量 d--螺栓螺纹公称直径 p--螺距 L--夹层厚度
螺纹伸出量
在螺栓与螺母连接的各种形式中，对于有锁紧(图1)或无琐紧(图2)的螺纹插入端(
图1
图2
螺纹在夹层中的相对位置
在拉力作用下工作的螺纹零件 螺纹部分(包括螺尾)在夹层中(壳体，零件)的长度
5.2 在剪力作用下工作的螺纹零件 5.2.1 光杆与夹层齐平 螺纹零件(螺栓)的光
图3
图4
5.2.2 螺纹在夹层外 螺纹零件(螺栓)的光杆部分占用夹层的全厚度时，即螺尾或
图5
```

Figure 4: A Side-by-Side Comparison of the Pairs of Text after Processing

2.2.20. Moses Server and Web Services

For the SCREAM Lab, a fully functioning system of translation that integrated the Moses Server for translation via web services was developed. The Web Services Description Language (WSDL) files were adapted to cooperate with the Foreign Media Collaboration Framework (FMCF) system described in Section 0.

Using simple access protocols and WSDL files, the methods for communicating with the Moses Server were adaptable across different systems. By following a set standard of method calls, various WSDL files were created for single-function operations. This was done to assure that the information the FMCF relied on for transmission of language pairs, port numbers, user id, priority status and source text would be initialized correctly.

The next step in the process was to create a single WSDL file that incorporated all of these functions and start testing them out with the FMCF system. Unfortunately Northrop Grumman lost their funding for the FMCF work and this action was stalled.

2.2.21. MT Eval Updates

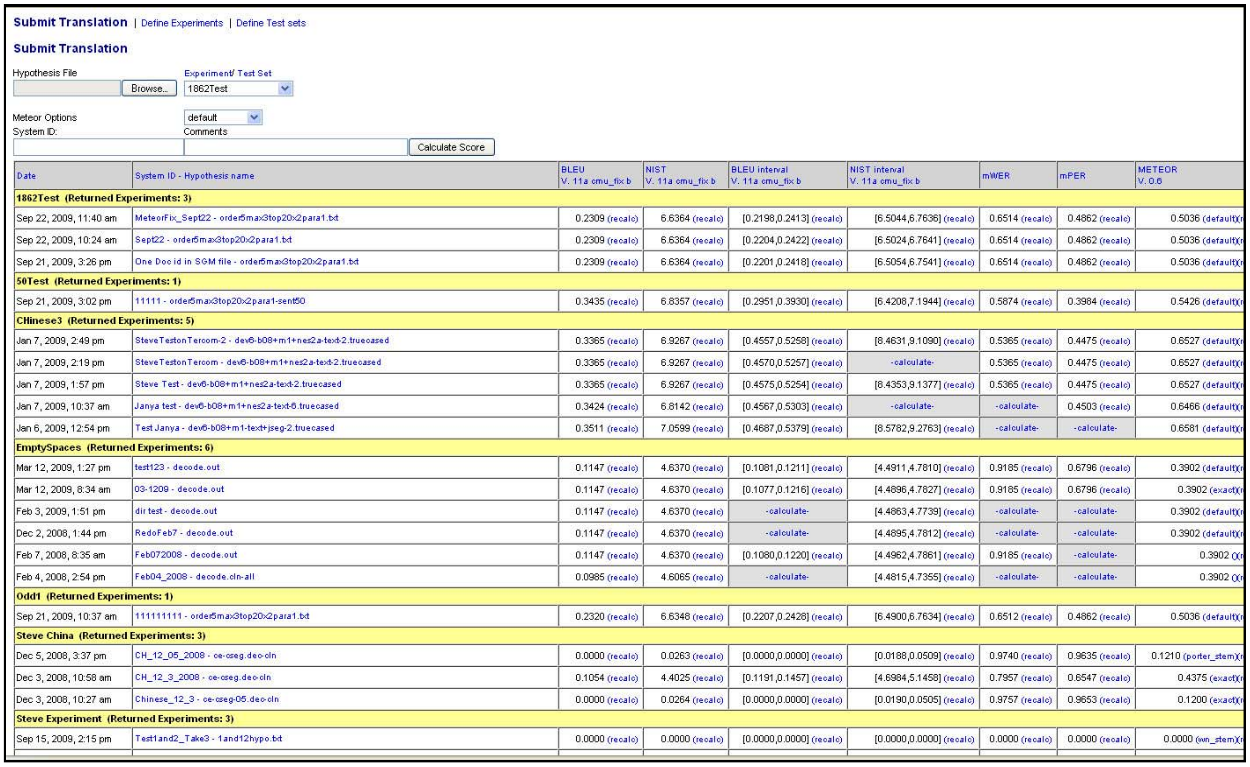
Carnegie Mellon University developed an online server for MT Evaluation (MT Eval) so that users could define their own test sets and experiments and have their scores automatically calculated. The interface was designed to ease the organization and archival of the score sets, but since its original design dates back to 2005, the SCREAM Lab added functionality to it as needed. The MT Eval translation score overview table is shown in Figure 5.

A user can define a new test set by choosing the reference text, source file and target/source languages. Experiments are created by choosing the text set, preprocessing steps (lower case, punctuation, etc) and which scores to calculate. Once the test set and experiment are created a user can then submit a translation hypothesis and have the scores calculated.

Over time, some scoring calculations were updated and the new calculations needed to be added to the server. The code was rewritten to find and initiate the new Java Jar files, test their output, and to stay knowledgeable of any new switches or functions included in the Jar file.

The MT Eval server was also updated to allow a means to view the translation hypothesis with its unique scoring on a line-by-line basis. PHP scripts were developed to take this output and capture it for Adobe PDF creation that could be viewed online or downloaded for offline study.

A third party Perl script which allows for color coding of n-gram text was ported over to PHP script and incorporated into the hypothesis viewer. Each unique color represented a unigram, bigram or trigram and could be turned off and on dynamically, as shown in Figure 6.



Submit Translation | Define Experiments | Define Test sets

Submit Translation

Hypothesis File: Browse...

Experiment/ Test Set: 1862Test

Meteor Options: default

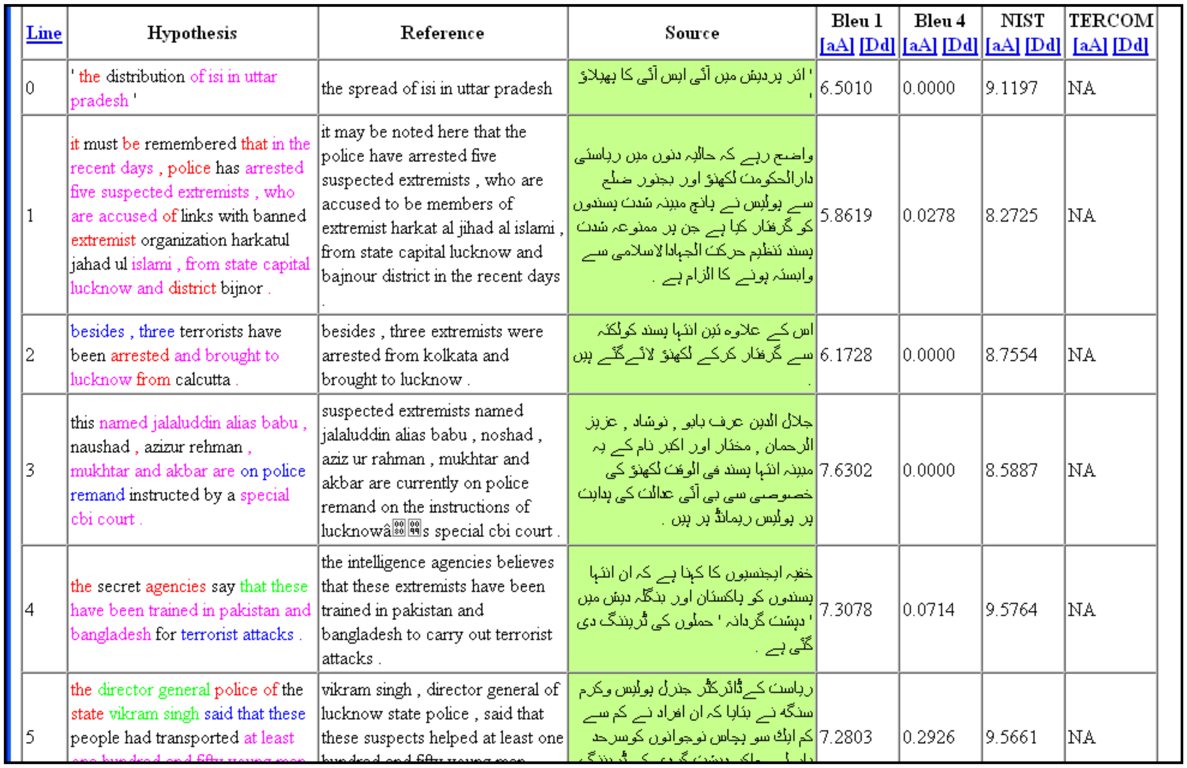
System ID:

Comments:

Calculate Score

Date	System ID - Hypothesis name	BLEU V. 11a cmu_fix b	NIST V. 11a cmu_fix b	BLEU interval V. 11a cmu_fix b	NIST interval V. 11a cmu_fix b	mWER	mPER	METEOR V. 0.6
1862Test (Returned Experiments: 3)								
Sep 22, 2009, 11:40 am	MeteorFix_Sept22 - order5max3top20x2para1.txt	0.2309 (recalc)	6.6364 (recalc)	[0.2198,0.2413] (recalc)	[6.5044,6.7636] (recalc)	0.6514 (recalc)	0.4862 (recalc)	0.5036 (default)(r
Sep 22, 2009, 10:24 am	Sept22 - order5max3top20x2para1.txt	0.2309 (recalc)	6.6364 (recalc)	[0.2204,0.2422] (recalc)	[6.5024,6.7641] (recalc)	0.6514 (recalc)	0.4862 (recalc)	0.5036 (default)(r
Sep 21, 2009, 3:26 pm	One Doc id in SGM file - order5max3top20x2para1.txt	0.2309 (recalc)	6.6364 (recalc)	[0.2201,0.2418] (recalc)	[6.5054,6.7541] (recalc)	0.6514 (recalc)	0.4862 (recalc)	0.5036 (default)(r
50Test (Returned Experiments: 1)								
Sep 21, 2009, 3:02 pm	11111 - order5max3top20x2para1-sent50	0.3435 (recalc)	6.8357 (recalc)	[0.2951,0.3930] (recalc)	[6.4208,7.1944] (recalc)	0.5874 (recalc)	0.3984 (recalc)	0.5426 (default)(r
CHinese3 (Returned Experiments: 5)								
Jan 7, 2009, 2:49 pm	SteveTestonTercom-2 - dev6-b08+m1+nes2a-text-2.truecased	0.3365 (recalc)	6.9267 (recalc)	[0.4557,0.5258] (recalc)	[8.4631,9.1090] (recalc)	0.5365 (recalc)	0.4475 (recalc)	0.6527 (default)(r
Jan 7, 2009, 2:19 pm	SteveTestonTercom - dev6-b08+m1+nes2a-text-2.truecased	0.3365 (recalc)	6.9267 (recalc)	[0.4570,0.5257] (recalc)	-calculate-	0.5365 (recalc)	0.4475 (recalc)	0.6527 (default)(r
Jan 7, 2009, 1:57 pm	Steve Test - dev6-b08+m1+nes2a-text-2.truecased	0.3365 (recalc)	6.9267 (recalc)	[0.4575,0.5254] (recalc)	[8.4353,9.1377] (recalc)	0.5365 (recalc)	0.4475 (recalc)	0.6527 (default)(r
Jan 7, 2009, 10:37 am	Janya test - dev6-b08+m1+nes2a-text-6.truecased	0.3424 (recalc)	6.8142 (recalc)	[0.4567,0.5303] (recalc)	-calculate-	-calculate-	0.4503 (recalc)	0.6466 (default)(r
Jan 6, 2009, 12:54 pm	Test Janya - dev6-b08+m1-text+jseg-2.truecased	0.3511 (recalc)	7.0599 (recalc)	[0.4687,0.5379] (recalc)	[8.5782,9.2763] (recalc)	-calculate-	-calculate-	0.6581 (default)(r
EmptySpaces (Returned Experiments: 6)								
Mar 12, 2009, 1:27 pm	test123 - decode.out	0.1147 (recalc)	4.6370 (recalc)	[0.1081,0.1211] (recalc)	[4.4911,4.7810] (recalc)	0.9185 (recalc)	0.6796 (recalc)	0.3902 (default)(r
Mar 12, 2009, 8:34 am	03-1209 - decode.out	0.1147 (recalc)	4.6370 (recalc)	[0.1077,0.1216] (recalc)	[4.4896,4.7827] (recalc)	0.9185 (recalc)	0.6796 (recalc)	0.3902 (exact)(r
Feb 3, 2009, 1:51 pm	dir test - decode.out	0.1147 (recalc)	4.6370 (recalc)	-calculate-	[4.4863,4.7739] (recalc)	-calculate-	-calculate-	0.3902 (default)(r
Dec 2, 2008, 1:44 pm	RedoFeb7 - decode.out	0.1147 (recalc)	4.6370 (recalc)	-calculate-	[4.4895,4.7812] (recalc)	-calculate-	-calculate-	0.3902 (default)(r
Feb 7, 2008, 8:35 am	Feb072008 - decode.out	0.1147 (recalc)	4.6370 (recalc)	[0.1080,0.1220] (recalc)	[4.4962,4.7861] (recalc)	0.9185 (recalc)	-calculate-	0.3902 ()(r
Feb 4, 2008, 2:54 pm	Feb04_2008 - decode.cln-all	0.0985 (recalc)	4.6065 (recalc)	-calculate-	[4.4815,4.7355] (recalc)	-calculate-	-calculate-	0.3902 ()(r
Odd1 (Returned Experiments: 1)								
Sep 21, 2009, 10:37 am	111111111 - order5max3top20x2para1.txt	0.2320 (recalc)	6.6348 (recalc)	[0.2207,0.2428] (recalc)	[6.4900,6.7634] (recalc)	0.6512 (recalc)	0.4862 (recalc)	0.5036 (default)(r
Steve China (Returned Experiments: 3)								
Dec 5, 2008, 3:37 pm	CH_12_05_2008 - ce-cseg.dec-cln	0.0000 (recalc)	0.0263 (recalc)	[0.0000,0.0000] (recalc)	[0.0188,0.0509] (recalc)	0.9740 (recalc)	0.9635 (recalc)	0.1210 (porter_stem)(r
Dec 3, 2008, 10:58 am	CH_12_3_2008 - ce-cseg.dec-cln	0.1054 (recalc)	4.4025 (recalc)	[0.1191,0.1457] (recalc)	[4.6984,5.1458] (recalc)	0.7957 (recalc)	0.6547 (recalc)	0.4375 (exact)(r
Dec 3, 2008, 10:27 am	Chinese_12_3 - ce-cseg-05.dec-cln	0.0000 (recalc)	0.0264 (recalc)	[0.0000,0.0000] (recalc)	[0.0190,0.0505] (recalc)	0.9757 (recalc)	0.9653 (recalc)	0.1200 (exact)(r
Steve Experiment (Returned Experiments: 3)								
Sep 15, 2009, 2:15 pm	Test1and2_Take3 - 1and12hypo.txt	0.0000 (recalc)	0.0000 (recalc)	[0.0000,0.0000] (recalc)	[0.0000,0.0000] (recalc)	0.0000 (recalc)	0.0000 (recalc)	0.0000 (wn_stem)(r

Figure 5: MT Eval Translation Score Overview Table



Line	Hypothesis	Reference	Source	Bleu 1 [aA] [Dd]	Bleu 4 [aA] [Dd]	NIST [aA] [Dd]	TERCOM [aA] [Dd]
0	' the distribution of isi in uttar pradesh '	the spread of isi in uttar pradesh	' اتر بردیش مین آئی ایس آئی کا بھیلاؤ '	6.5010	0.0000	9.1197	NA
1	it must be remembered that in the recent days , police has arrested five suspected extremists , who are accused of links with banned extremist organization harkatul jahad ul islami , from state capital lucknow and district bijnor .	it may be noted here that the police have arrested five suspected extremists , who are accused to be members of extremist harkat al jihad al islami , from state capital lucknow and bajnour district in the recent days .	واضح رہے کہ حالیہ دنوں مین ریاستی دارالحکومت لکھنؤ اور بجنور ضلع سے بولیس نے بانچ مبینہ شدت بسندوں کو گرفتار کیا ہے جن بر ممنوعہ شدت بسند تنظیم حرکت الجہادالاسلامی سے وابستہ ہونے کا الزام ہے .	5.8619	0.0278	8.2725	NA
2	besides , three terrorists have been arrested and brought to lucknow from calcutta .	besides , three extremists were arrested from kolkata and brought to lucknow .	اس کے علاوہ تین انتہا بسند کولکتہ سے گرفتار کرکے لکھنؤ لائےگئے ہین .	6.1728	0.0000	8.7554	NA
3	this named jalaluddin alias babu , naushad , azizur rehman , mukhtar and akbar are on police remand instructed by a special cbi court .	suspected extremists named jalaluddin alias babu , noshad , aziz ur rahman , mukhtar and akbar are currently on police remand on the instructions of lucknowâ��s special cbi court .	جلال الدین عرف بابو , نوشاد , عزیز الرحمان , مختار اور اکبر نام کے یہ مبینہ انتہا بسند فی الوقت لکھنؤ کی خصوصی سی بی آئی عدالت کی ہدایت بر بولیس ریمانڈ بر ہین .	7.6302	0.0000	8.5887	NA
4	the secret agencies say that these have been trained in pakistan and bangladesh for terrorist attacks .	the intelligence agencies believes that these extremists have been trained in pakistan and bangladesh to carry out terrorist attacks .	خفیہ ایجنسیوں کا کہنا ہے کہ ان انتہا بسندوں کو باکستان اور بنگلہ دیش مین ' دہشت گردانہ ' حملوں کی ٹریننگ دی گئی ہے .	7.3078	0.0714	9.5764	NA
5	the director general police of the state vikram singh said that these people had transported at least one hundred and fifty young men	vikram singh , director general of lucknow state police , said that these suspects helped at least one hundred and fifty young men	ریاست کے ڈائرکٹر جنرل بولیس وکرم سنگھ نے بتایا کہ ان افراد نے کم سے کم ایک سو بچاس نوجوانوں کوسرحد	7.2803	0.2926	9.5661	NA

Figure 6: MT Eval Hypothesis Viewer with Active Ngram Color Coding

2.2.22. Summary

Various improvements were developed for the SMT2 SMT system. Character level processing was improved by reducing variation with spelling normalization and tokenization. Morphological processing was used to improve recognition different instances of the same word. Errors in sentence alignment were detected and corrected. Word alignment errors were analyzed to reduce phrase table errors. Several paraphrasing software programs were explored to generate variant sentences with similar meanings for use in training or scoring of MT.

Performance improvements, functionality enhancements, and error corrections were performed on various software packages which are actively being used in the SCREAM Laboratory.

Scripts were written to convert XML-based TW files into parallel text for LM training. Web service files were created to begin communications between the Moses Translator and the FMCF server. The Carnegie Mellon MT Eval server had various scoring displays and functions added to its framework.

2.2.23. Recommendations for Future Work

Spelling normalization for Dari should be investigated further, since there are variant characters (Farsi yeh 06CC ی and Arabic yeh 064A ي) that occur in similar proportions in the data, according to the preference of the writer.

Pashto machine translation might be improved by creating a Pashto morphological analyzer, possibly built within the GF/Molto rule system, using published morphological rules for Pashto noun and verb inflection.

Lattice and confusion net representations were explored for paraphrasing; lattices could also be applied to represent other variations, including punctuation variants (e.g., French « » vs. "), inflected forms, and semantic alternatives such as the simple vs. compound past forms of French verbs.

Current punctuation normalization and tokenization methods depend on well-defined lists of punctuation characters. However, data across languages show unexpected use of punctuation, such as the use of hyphens in place of the Urdu full stop character ۔ 06D4. These substitutions can be driven by visual similarity and by limitations of the input methods, and this type of variation increases in informal text such as blogs and social media. Future work should therefore consider more dynamic methods of recognizing and normalizing punctuation variation.

Sentence alignment corrections did not provide the expected benefits for translation. Instead, further research should be focus on word alignment and its relationship to phrase table extraction, including the use of word alignment weights and the possible use of validated word alignments to delimit phrases.

Performance enhancements should be continually researched, using both traditional techniques and utilizing a GPGPU. Additionally, a 'translation memory' package may be beneficial and should be evaluated, as described in "Convergence of Translation Memory and Statistical Machine Translation" [15].

Finally, the MT Eval server functionality could be embedded within the Experiment Reader software in the SCREAM Lab, allowing for more control of sorting and scoring MT output.

2.3 Speech Synthesis

This section discusses the speech synthesis systems that were developed. Section 0 describes three modifications that were made to the HMM-based Speech Synthesis System (HTS) to reduce the model training time. Section 0 describes three English HTS systems that were developed using full context labels. Section 0 presents a Graphical User Interface (GUI) that was developed for rapidly adapting an existing speech synthesis system to fit the characteristics of a new speaker. Finally, Section 0 summarizes the speech synthesis experiments and Section 0 provides recommendations for future work.

2.3.1. HTS Modifications

This section discusses modifications that were made to HTS.[12] First, configuration settings were added to specify whether an HTS program should write an entire HMM set, only the physical HMM macros,[13] or all macros except for the physical HMM macros. This is useful because the physical HMM macros do not change after decision tree clustering.

Next, the code was modified to optimize the accumulation of statistics when estimating feature transforms. As mentioned in Section 0, feature transforms are used to reduce the mismatch between an AM and a set of feature vectors. When using class-based transforms, Gaussian mixture components are grouped into different classes and a transform is estimated for each class. In ASR systems it is common to group the Gaussian components into a relatively small number of classes. For example, the systems discussed in Section 0 grouped the Gaussian components into 32 different classes and each class included up to 8649 Gaussians. In HMM-based speech synthesis systems, however, it is not uncommon to use hundreds of thousands of classes and for each class to include only a few Gaussian components. For example, the English HTS system described later in this section included 223,890 classes and each class included less than 4 Gaussians. It was discovered that when processing each feature vector, the HTS code loops through all classes to update the accumulator statistics. This makes sense for ASR systems because there are only a few classes and it is common to have non-zero accumulator updates for several classes after processing each feature vector. In HTS systems, however, the accumulator updates are zero for the majority of the classes. The code was modified to keep track of which classes need to be processed when updating the accumulator statistics.

Finally, support was added to the HERest program to iteratively estimate feature transforms. This is useful for reducing the overhead associated with reading and writing the HMM set between iterations. These modifications were evaluated on an HMM speaker adaptation task. The average voice model was trained 39 hours of speech produced by 163 male speakers from the Wall Street Journal (WSJ) corpus [16, 17]. The models were developed using the same training procedure and default configuration settings distributed with the HTS speaker adaption demo.[14] The final HMM set included 1,342,104 physical HMMs and a total of 223,890 classes were used for adaptation. Table 3 shows the computation time required to adapt the average voice model using 40 utterances. Note that the *optimized model writing* in Table 3 included the

[12] Available at: http://hts.sp.nitech.ac.jp

[13] The ~h macro described in the HTK Book [2]

[14] Available at: http://hts.sp.nitech.ac.jp/archives/2.1.1/HTS-demo_CMU-ARCTIC-ADAPT.tar.bz2

Table 3: Time Required to Adapt an HTS Speech Synthesis System

Three different modifications were made to the HTS programs and training script.

Modifications	Hours
None	19.7
Optimized model writing	4.1
+ Optimized transform estimation	1.0
+ HERest iterative transform estimation	0.7

following changes: the physical HMM definitions were not written when updating the models, and the HMM list was not compacted to merge HMMs that shared the same distributions.[15]

2.3.2. English Full Context Models

HMM-based speech synthesis systems were developed for English using full context labels. In ASR systems, HMMs are typically trained to model triphones. HMM-based speech synthesis systems can also be trained to model triphones, although more natural speech can be synthesized by incorporating additional contextual information. The full context labels used in this experiment are the same as distributed with the HTS speaker adaption demo, and include contextual factors such as syllable stress, syllable identity, and word position.

Full context labels were generated for the WSJ corpus using Festival,[16] and initial time alignments were produced using an HTK ASR system trained on the same data set. Average voice models were developed on three subsets of the WSJ corpus using the same training procedure and default configuration settings distributed with the HTS speaker adaption demo. Each subset included the same 163 male speakers and was created by selecting a maximum of N utterances per speaker. The first model was trained on 12 hours of speech selected using $N = 50$; the second model was trained on 23 hours of speech selected using $N = 100$; and the third model was trained on 39 hours of speech selected using $N = 200$. No formal listening tests were conducted to compare the models, although it is the author's opinion that increasing the amount of speech data improved the overall voice quality.

2.3.3. Speaker Adaptation GUI

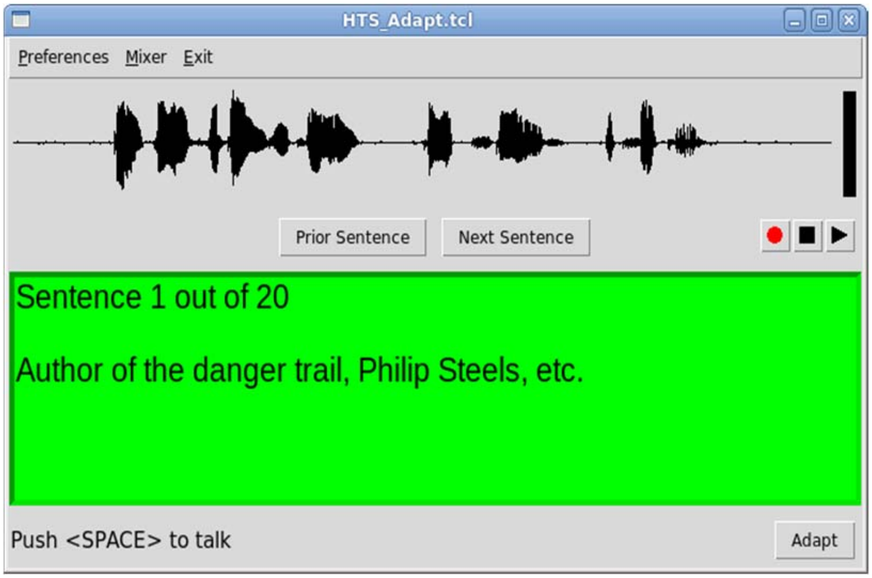
A GUI was developed for recording speech and adapting an HMM-based speech synthesis system.[17] Figure 7 shows a screenshot of the GUI. The user is first prompted to record a set of sentences. Audio playback and recording are provided using the Snack Sound Toolkit.[18] After recording the sentences, the following procedure is applied to adapt the HMMs:

[15] The HHEd CO command described in the HTK Book [2]

[16] Available at: http://www.cstr.ed.ac.uk/projects/festival

[17] The GUI is a modified version of the SCREAM Lab Recorder developed by Mr. Eric Hansen

[18] Available at: http://www.speech.kth.se/snack



Sentence 1 out of 20

Author of the danger trail, Philip Steels, etc.

Push <SPACE> to talk

Figure 7: Screenshot of the GUI Used to Record Sentences and Adapt an HMM-Based Speech Synthesis System

Generate full context labels for each sentence using Festival

1. Compute spectrum and pitch features for each audio file
2. Synthesize unseen non-SAT HMMs using the decision tree questions
3. Adapt the non-SAT HMMs
4. Synthesize unseen SAT HMMs using the decision tree questions
5. Adapt the SAT HMMs using the models estimated in Step 4 for the initial alignment
6. Convert the models to hts_engine format

The configuration settings are specified using the same file format as distributed with the HTS speaker adaption demo. To reduce the computation time, an option was added for skipping Steps 3–4 and using the SAT HMMs for the initial alignment in Step 6. In the author's opinion, this has a negligible effect on the final voice quality.

2.3.4. HTK Improvements

The HTK code was modified to support optional compilation for 64-bit platforms, including HDecode. Compile-time checking of the size of a pointer is used to determine 32 versus 64 bit compilation. It was also determined that using the Intel C++ Compiler can result in an up to 30% increase in performance.

The feasibility of porting HTK to use the GPGPU was researched and found to not be suitable without a comprehensive redesign.

2.3.5. Summary

The HTS code was modified to reduce the overall computation time. First, the model writing procedure was optimized so that the physical HMM definitions are only written when required and the HMM list is not compacted. Next, the code was modified to optimize the accumulation of statistics when estimating feature transforms. Finally, support was added to the HERest program to iteratively estimate feature transforms. These changes reduced the overall computation time from 19.7 hours to 0.7 hours on an HMM speaker adaptation task.

Three English HMM-based speech synthesis systems were developed using full context labels. The models were trained on 12 hours, 23 hours, and 39 hours of speech spoken by 163 male speakers from the WSJ corpus. Lastly, a GUI was developed for recording speech and adapting an HMM-based speech synthesis system.

2.3.6. Recommendations for Future Work

English HMM-based speech synthesis systems were only developed using male speakers; it would be useful to train systems on female speakers as well. A common method for developing an HMM-based speech synthesis system is to train an average voice model from multiple speakers and then adapt the model to fit the characteristics of a single speaker. New voices could be created by adapting the models described in Section 0. In addition, it could be beneficial to conduct a formal listening test to compare the systems.

The GUI used to record speech currently relies on the Snack Sound Toolkit. The most recent release of this toolkit from the Royal Institute of Technology (KTH) was in 2005, and on our newest computers we have experienced problems interfacing with the sound card. It would be worthwhile to implement an alternative method for recoding and playing audio.

2.4 Spoken Language Translation Systems / Text Translation Systems

This section discusses Spoken Language Translation Systems and Text Translation Systems. Section 0 covers Prototype Systems Development, section 0 covers Operational Systems Development, and sections 0 and 0 provide a summary and recommendations for future work.

2.4.1. Prototype Systems Development

The FMCF developed by Northrop Grumman is a Service Oriented Architecture (SOA) that allows for MT, optical character recognition (OCR) and ASR through a set of collaboration tools.

Northrop Grumman has been developing an incarnation of the FMCF for use by NASIC and wanted to incorporate the Moses Server translation web service into the toolset.

Northrop Grumman allowed the SCREAM Lab to use a version of FMCF on a laptop that was partitioned off so that a Linux installation could be installed with PHP capabilities, allowing communication with FMCF.

Once PHP was installed on the laptop work began on installing the script for basic communication with the SCREAM network and the development of the WSDL files for translation through the Moses Server.

The documentation provided for FMCF's own web services described the method calls it used for file queuing, translation and retrieval. The WSDL services would need to be able to communicate from the Moses Servers' specific translation ports and send it back to FMCF.

A series of basic WSDL files were created to define the method calls and to check the validity of the web services process. The set of WSDL files were then culled into one main WSDL file to support testing on the SCREAM network. A portion of this file can be seen in Figure 8. Ideally, FMCF would have been updated to allow for the calls to the Moses Server WSDL.

Regrettably, Northrop Grumman lost their financing through NASIC and the work remains incomplete on the laptop. Since that time some of the licenses to FMCF collaboration tools have expired and it is impossible to even continue on this endeavor.

2.4.2. Operational Systems Development

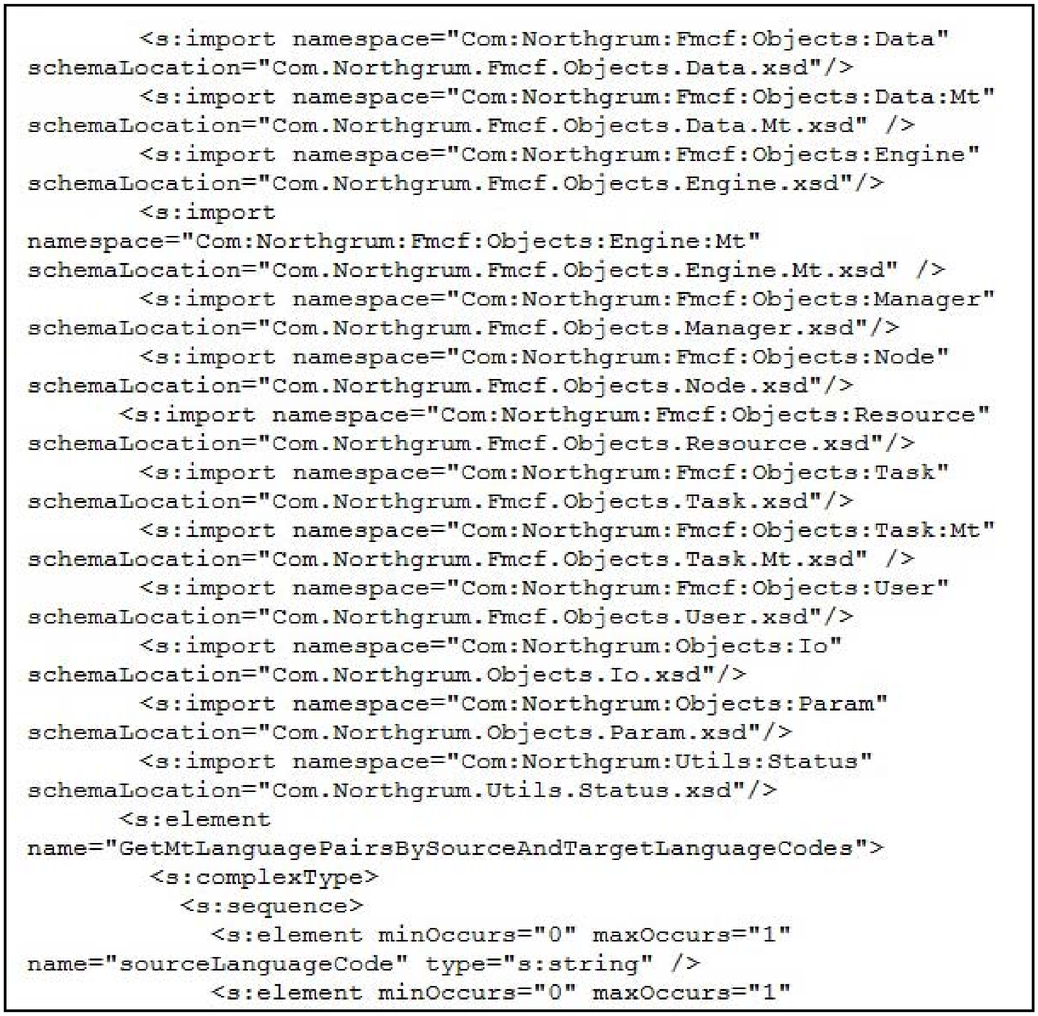
Due to the problems described at the end of Section 0, work on operational prototype SLTS and TTS systems could not proceed.

2.4.3. Summary

PHP and MySQL were installed on the FMCF Laptop to begin WSDL communication between the laptop and the SCREAM network. Because of lack of funding for the Northrop Grumman work, this work has stalled.

2.4.4. Recommendations for Future Work

It is recommended that there is continued development of web services for translations between separate systems based on the knowledge learned through the initial steps of binding to the FMCF server.



```
        <s:import namespace="Com:Northgrum:Fmcf:Objects:Data"
schemaLocation="Com.Northgrum.Fmcf.Objects.Data.xsd"/>
        <s:import namespace="Com:Northgrum:Fmcf:Objects:Data:Mt"
schemaLocation="Com.Northgrum.Fmcf.Objects.Data.Mt.xsd" />
        <s:import namespace="Com:Northgrum:Fmcf:Objects:Engine"
schemaLocation="Com.Northgrum.Fmcf.Objects.Engine.xsd"/>
        <s:import
namespace="Com:Northgrum:Fmcf:Objects:Engine:Mt"
schemaLocation="Com.Northgrum.Fmcf.Objects.Engine.Mt.xsd" />
        <s:import namespace="Com:Northgrum:Fmcf:Objects:Manager"
schemaLocation="Com.Northgrum.Fmcf.Objects.Manager.xsd"/>
        <s:import namespace="Com:Northgrum:Fmcf:Objects:Node"
schemaLocation="Com.Northgrum.Fmcf.Objects.Node.xsd"/>
       <s:import namespace="Com:Northgrum:Fmcf:Objects:Resource"
schemaLocation="Com.Northgrum.Fmcf.Objects.Resource.xsd"/>
        <s:import namespace="Com:Northgrum:Fmcf:Objects:Task"
schemaLocation="Com.Northgrum.Fmcf.Objects.Task.xsd"/>
        <s:import namespace="Com:Northgrum:Fmcf:Objects:Task:Mt"
schemaLocation="Com.Northgrum.Fmcf.Objects.Task.Mt.xsd" />
        <s:import namespace="Com:Northgrum:Fmcf:Objects:User"
schemaLocation="Com.Northgrum.Fmcf.Objects.User.xsd"/>
        <s:import namespace="Com:Northgrum:Objects:Io"
schemaLocation="Com.Northgrum.Objects.Io.xsd"/>
        <s:import namespace="Com:Northgrum:Objects:Param"
schemaLocation="Com.Northgrum.Objects.Param.xsd"/>
        <s:import namespace="Com:Northgrum:Utils:Status"
schemaLocation="Com.Northgrum.Utils.Status.xsd"/>
      <s:element
name="GetMtLanguagePairsBySourceAndTargetLanguageCodes">
        <s:complexType>
          <s:sequence>
            <s:element minOccurs="0" maxOccurs="1"
name="sourceLanguageCode" type="s:string" />
            <s:element minOccurs="0" maxOccurs="1"
```

Figure 8: WSDL for Communication between FMCF and SCREAM Lab

2.5 Laboratory Corpora Support

Corpora were created or extended for various languages, including French, Dari, Pashto, Arabic, Hindi, Urdu, and Hausa; corpus resources were identified for Farsi, Swahili, Somali, Hausa, Igbo, and Yoruba. The Global Language Online Support System (GLOSS) website was used to generate parallel text. Section 0 discusses parallel text, Section 0 covers the identification of Lexical resources, Section 0 mentions treebank resources, Section 0 identifies other MT resources, and Section 0 discusses parallel text resources via human translators.

2.5.1. Parallel Text

The development of good MT models requires large amounts of parallel text from which the system can extract translation patterns. This section reports the identification of online resources for parallel text documents for Dari, Pashto, Farsi, French, Swahili, Hausa, Somali, Yoruba, and Igbo, as well as parallel text from online language classes for Arabic, Dari, Hindi, Pashto, Hausa, and Urdu. Also reported here are the steps taken to verify that extracted text has been assigned to the correct language, and that text derived from PDFs has been extracted in the correct order.

Identification of Parallel Text Resources Online:

Websites can provide parallel text for certain language pairs. Identifying these websites involves searching for the language name or typical words of the language, and examining sample text to see if the correct language is present and if the material in the two languages is actually parallel.

English/Dari/Pashto: Websites containing parallel text in English, Dari, and Pashto include the previously identified Sada-e-Azadi news website, various Afghanistan government websites, and the Institute for War and Peace Reporting website. (see Appendix A)

English/Farsi: Parallel text in English and Farsi can be used to help translate the English/Dari language pair. Websites containing English and Farsi text include: the US Virtual Embassy to Iran, the US State Department Bureau of International Information Programs, and the Central Asia Online website (with English, Farsi, Urdu, and Russian). (see Appendix B)

African Languages: Swahili, Somali, Hausa, Igbo, Yoruba:

An initial search was conducted to identify online resources for African languages, including Swahili, Somali, Hausa, Igbo, and Yoruba. Health and legal documents for refugee populations were the primary sources identified for text with English translations. (see Appendix C)

French/English TED Talks: The TED talk dataset includes both parallel text, with matching articles in French and English, and monolingual text, with articles just in English. This monolingual data was supplemented by identifying the corresponding French data from a webscrape. Programs were written to format the data (removing timestamps, combining sentence fragments, and adding XML tags). The resulting files were hand-edited to ensure parallel sentence breaks across the French and English files.

French Hand-Aligned Data: Two sources of French data were identified in which the word alignments have been hand-edited by human translators. These are the Hansard hand-aligned corpus[19] and the Europarl hand-aligned corpus.[20] These contain 447 sentences of the Hansards of the Canadian parliament [18] and 100 sentences from the proceedings of the European Parliament [19], respectively.

Global Language Online Support System (GLOSS) Collection: The GLOSS website[21] is the Defense Language Institute's Foreign Language Center for independent study of foreign languages.

The GLOSS website was scraped to gather its parallel text and corresponding media files. A basic wget script was written to crawl through specific language directory structures and download it into a mirror of the site locally. This scraping was able to acquire the following courses:

[19] http://www.cs.unt.edu/~rada/wpt

[20] http://www.l2f.inesc-id.pt/resources/translation/

[21] http://gloss.dliflc.edu/Default.aspx

Table 4: Languages and Number of Courses from the GLOSS Website

Language:	Courses:
Arabic	194
Dari	156
Hausa	57
Hindi	150
Pashto	97
Urdu	50

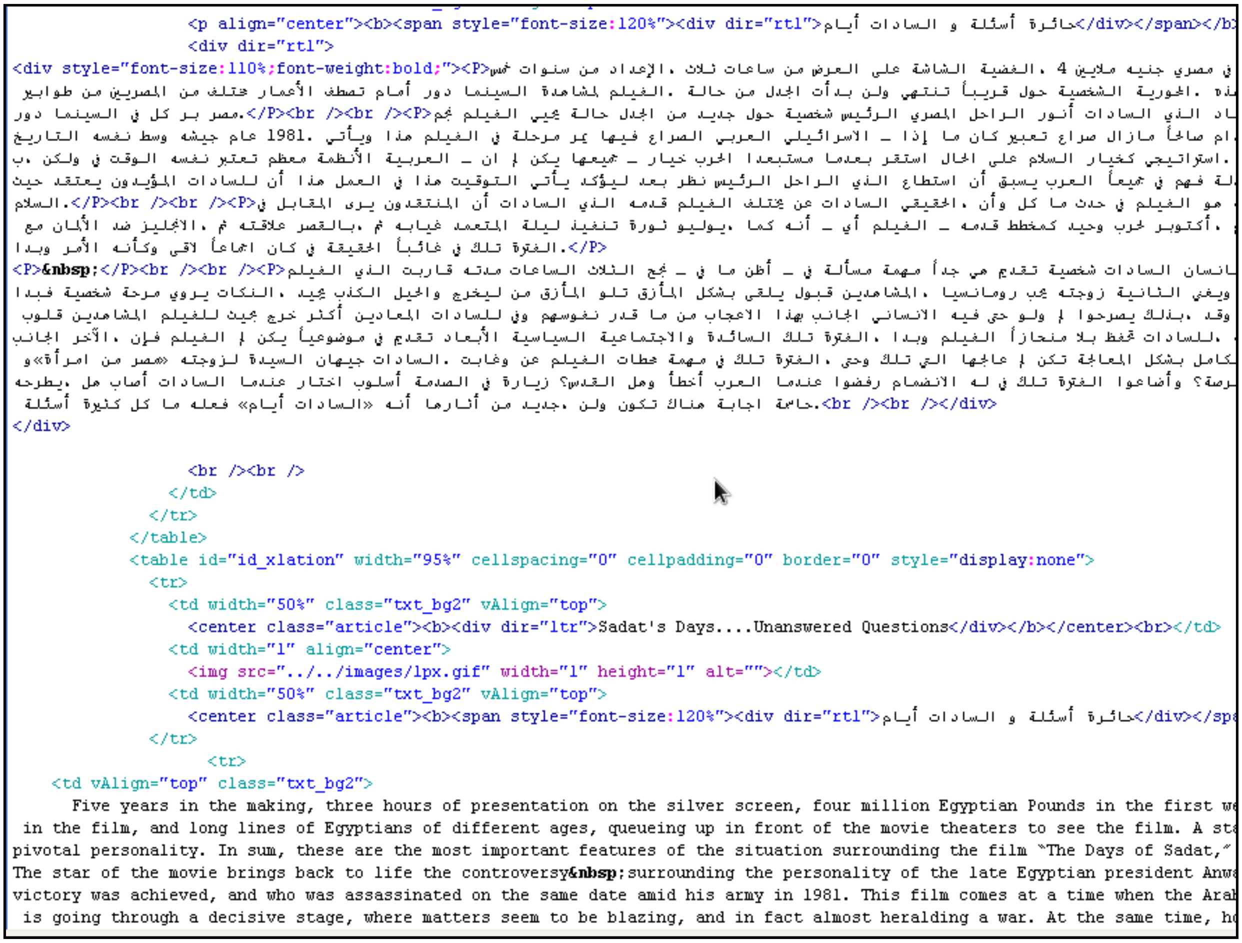
A Perl script was developed to parse the HTML and create separate text files for the language, English text, skill level and comprehension level. Table 4 shows the HTML of the GLOSS course webpage and Figure 10 shows the English and source language after separation. As of this report only the Arabic course files have been fully processed.

Language Identification: MT can be improved by identifying errors in which certain sections of the training data are in the wrong language. This language identification process must exclude direct quotations or named entities, however. An existing program, TextCat, was used for language identification based on ngram frequency models; specialized programs were also written to detect the language based on lexical coverage and character ranges, without the need to compute a frequency model.

Dari and Pashto Language Identification in the Sada-e-Azadi Data: Text extracted from the Sada-e-Azadi news website sometimes has errors in which an article is misidentified as Dari vs. Pashto or vice versa. To address this problem, a program was written to check individual words against a dictionary and report the percent of unknown words. Dari files should have fewer unknowns with the Dari lexicon, and Pashto files should have fewer unknowns with the Pashto lexicon.

A second program was written to distinguish Dari and Pashto text, based on the proportion of Pashto-specific characters (څ ځ ټ ډ ړ ښ ږ ګ ڼ ې ۍ), as well as the occurrence of the Pashto word د /d/ 'of'. In the data from Sada-e-Azadi, Pashto files typically have 75% or more lines identified as specifically Pashto. A Dari file that contains borrowed Pashto words will have some lines identified as Pashto, but the percentage is typically below 40%. Another indicator that could be used is the fact that Dari writers typically use Arabic numerals (123), while Pashto writers use Eastern Arabic numerals (١ ٢ ٣).

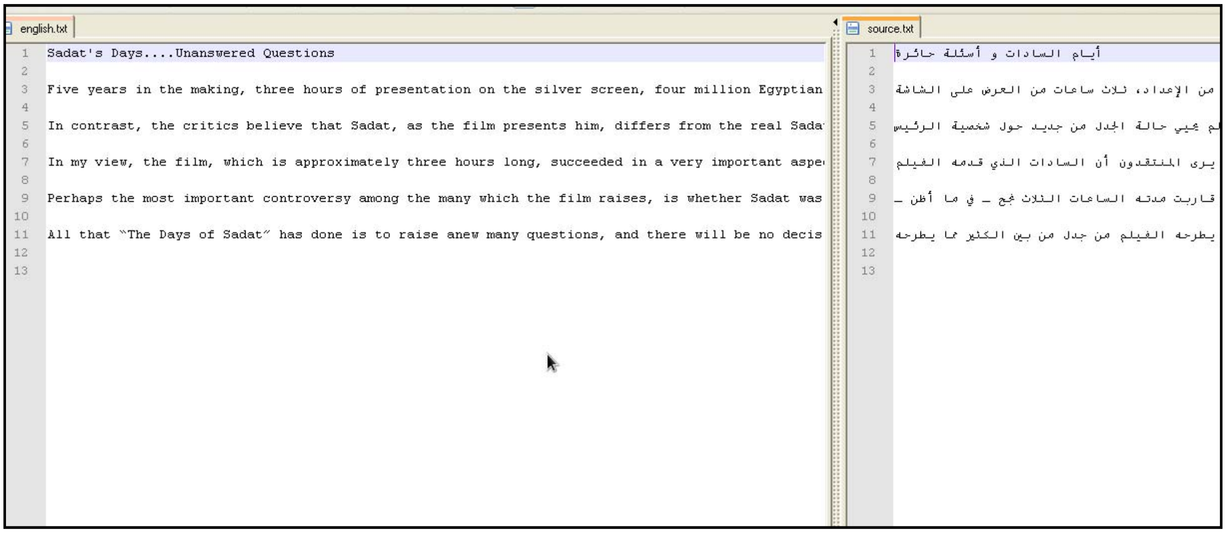
A program was written to detect Dari or Pashto characters within an English text, for use in troubleshooting parallel text. The program reports the proportion of non-English text, since a small amount could represent a legitimate use such as a quotation or a name spelled out in two languages.



```
                  <p align="center"><b><span style="font-size:120%"><div dir="rtl">أيام السادات و أسئلة حائرة</div></span></b
                  <div dir="rtl">
<div style="font-size:110%;font-weight:bold;"><P>خمس سنوات من الإعداد، ثلاث ساعات من العرض على الشاشة الفضية، 4 ملايين جنيه مصري في
 طوابير من المصريين من مختلف الأعمار تصطف أمام دور السينما لمشاهدة الفيلم. حالة من الجدل بدأت ولن تنتهي قريباً حول الشخصية المحورية. هذه
دور السينما في كل بر مصر.</P><br /><br /><P>نجم الفيلم يحيي حالة الجدل من جديد حول شخصية الرئيس المصري الراحل أنور السادات الذي اد
التاريخ نفسه وسط جيشه عام 1981. ويأتي هذا الفيلم في مرحلة يمر فيها الصراع العربي الاسرائيلي – إذا ما كان تعبير صراع مازال صالحاً ام
ب، ولكن في الوقت نفسه تعتبر معظم الأنظمة العربية – ان لم يكن جميعها – خيار الحرب مستبعدا بعدما استقر الحال على السلام كخيار استراتيجي.
حيث يعتقد المؤيدون للسادات أن هذا العمل في هذا التوقيت يأتي ليؤكد بعد نظر الرئيس الراحل الذي استطاع أن يسبق العرب جميعاً في فهم لة
السلام.</P><br /><br /><P>في المقابل يرى المنتقدون أن السادات الذي قدمه الفيلم يختلف عن السادات الحقيقي، وأن كل ما حدث في الفيلم هو
 مع الألمان ضد الانجليز، ثم علاقته بالقصر، ثم غيابه المتعمد ليلة تنفيذ ثورة يوليو، كما أنه – أي الفيلم – قدمه كمخطط وحيد لحرب أكتوبر،
وبدا الأمر وكأنه لاقى اجماعاً كان في الحقيقة غائباً في تلك الفترة.</P>
<P>&nbsp;</P><br /><br /><P>الفيلم الذي قاربت مدته الساعات الثلاث نجح – في ما أظن – في مسألة مهمة جداً هي تقديم شخصية السادات انسان
فبدا شخصية مرحة يروي النكات، يجيد الكذب والحيل ليخرج من المأزق تلو المأزق بشكل يلقى قبول المشاهدين، رومانسيا يحب زوجته الثانية ويغني
 قلوب المشاهدين للفيلم بحيث خرج أكثر المعادين للسادات وفي نفوسهم قدر ما من الاعجاب بهذا الجانب الانساني فيه حتى ولو لم يصرحوا بذلك، وقد
الجانب الآخر، فإن الفيلم لم يكن موضوعياً في تقدم الأبعاد السياسية والاجتماعية السائدة تلك الفترة، وبدا الفيلم منحازاً بلا تحفظ للسادات،
و «امرأة من مصر» لزوجته السيدة جيهان السادات. وغابت عن الفيلم محطات مهمة في تلك الفترة، وحتى تلك التي عالجها لم تكن المعالجة بشكل كامل
يطرحه، هل أصاب السادات عندما اختار أسلوب الصدمة في زيارة القدس؟ وهل أخطأ العرب عندما رفضوا الانضمام له في تلك الفترة وأضاعوا رصة؟
 أسئلة كثيرة كل ما فعله «أيام السادات» أنه أثارها من جديد، ولن تكون هناك اجابة حاسمة.<br /><br /></div>
</div>

                  <br /><br />
                </td>
              </tr>
            </table>
            <table id="id_xlation" width="95%" cellspacing="0" cellpadding="0" border="0" style="display:none">
              <tr>
                <td width="50%" class="txt_bg2" vAlign="top">
                  <center class="article"><b><div dir="ltr">Sadat's Days....Unanswered Questions</div></b></center><br></td>
                <td width="1" align="center">
                  <img src="../../images/1px.gif" width="1" height="1" alt=""></td>
                <td width="50%" class="txt_bg2" vAlign="top">
                  <center class="article"><b><span style="font-size:120%"><div dir="rtl">أيام السادات و أسئلة حائرة</div></spa
              </tr>
                    <tr>
        <td vAlign="top" class="txt_bg2">
          Five years in the making, three hours of presentation on the silver screen, four million Egyptian Pounds in the first we
 in the film, and long lines of Egyptians of different ages, queueing up in front of the movie theaters to see the film. A sta
pivotal personality. In sum, these are the most important features of the situation surrounding the film "The Days of Sadat,"
The star of the movie brings back to life the controversy&nbsp;surrounding the personality of the late Egyptian president Anwa
victory was achieved, and who was assassinated on the same date amid his army in 1981. This film comes at a time when the Arab
 is going through a decisive stage, where matters seem to be blazing, and in fact almost heralding a war. At the same time, ho
```

Figure 9: Embedded Languages in the HTML of the GLOSS Courses



english.txt

```
Sadat's Days....Unanswered Questions

Five years in the making, three hours of presentation on the silver screen, four million Egyptian

In contrast, the critics believe that Sadat, as the film presents him, differs from the real Sada

In my view, the film, which is approximately three hours long, succeeded in a very important aspe

Perhaps the most important controversy among the many which the film raises, is whether Sadat was

All that "The Days of Sadat" has done is to raise anew many questions, and there will be no decis
```

source.txt

```
أيام السادات و أسئلة حائرة

من الإعداد، ثلاث ساعات من العرض على الشاشة

لم يحيي حالة الجدل من جديد حول شخصية الرئيس

يرى المنتقدون أن السادات الذي قدمه الفيلم

قاربت مدته الساعات الثلاث نجح – في ما أظن –

يطرحه الفيلم من جدل من بين الكثير مما يطرحه
```

Figure 10: English and Source Files Created from Extraction of the GLOSS Course Files

Non-French Languages in the French Data: The French dataset contains some sections of wrong-language text. The newscom articles include English and German in the French; the Europarl articles include English, Greek, Russian, and other languages in the French; and the gigaword French files also include Canadian languages like Inuktitut.

A program was written using character ranges to find sections of non-Latin characters in the French data. This enabled the removal of sections of Greek and Russian text, for example.

It is harder to detect English in the French, since both languages use the Latin alphabet, and the languages have many identical words (e.g., *identification*). A program was written to report sentences that match exactly across the English and the French files. These are sometimes legitimate, involving either matching English and French words, names, or direct quotations. Longer identical sentences are likely to be errors, however. A second program was written that uses the Levenshtein edit distance to detect parts of sentences that have been copied instead of translated.

PDF Extraction Reversals: When text is extracted from the Sada-e-Azadi PDF files, about half the files have the letters reversed (even though the letters display correctly in the original PDF file). Reversed files were detected by calculating the percentage of unknown words using a language-specific dictionary. For the Sada-e-Azadi data, actual Dari files tend to have 60% or more words found in the current Farsi lexicon, while actual Pashto files tend to have 80% or more words found in the current Pashto lexicon. Files below these levels were candidates for a second program that reverses each line. The reversed files were then run through the dictionary check again; a decrease in unknown words shows that the reversal has restored the correct order.

Unfortunately, the letter-reversal program is insufficient to restore the reversed articles. The column format used in the Sada-e-Azadi articles causes the text to be extracted in sections, and therefore reversed text needs to be re-constituted section by section. After extraction, however, the column break information is no longer available.

Diagram of Column Reversal Problem

PDF Display	**Extraction Error**	**After Letter Reversal**	**Required Order**
FE DC BA ML KJ IH	FE DC BA ML KJ IH	HI JK LM AB CD EF	AB CD EF HI JK LM

Additional problems arise when the text contains digits, or words in Latin characters, which both run left to right within the right-to-left Arabic script document.

A second method was developed to correct the reversed files, using the original locational data for each character. The existing program PDFMiner[22] is used to generate a list of characters and xy-coordinates; a new program, pdfminer2text.pl, was written to re-create the words in order. This program detects the presence of Arabic or Latin characters or digits, and outputs the appropriate right-to-left or left-to-right order.

2.5.2. Lexica

This section reports the identification of online lexical resources for Dari, Pashto, Urdu, Swahili, Hausa, Somali, Yoruba, Igbo, and French, as well as the improvement of existing lexical resources for Pashto and Urdu.

[22] http://www.unixuser.org/~euske/python/pdfminer/

Identified Lexicon Resources Online: A search was conducted for online lexicon resources for various languages. For French, only a few small dictionaries were freely available; larger, older dictionaries are available only by purchase. For the African languages, online dictionaries were located of varying size and quality for Swahili, Hausa, Somali, Yoruba, and Igbo. (see Appendix C)

The website for the Digital Dictionaries of South Asia (DSAL) was identified as a source of lexicon resources for Dari, Pashto, and Urdu. This includes the Raverty dictionary for Pashto, which contains almost 15,000 entries, and the smaller Heston dictionary, with about 1300 entries. (see Appendix A) The format of these dictionaries makes it difficult to extract the entries; so far, headwords have been extracted from the Raverty dictionary.

Online flashcards were identified as a supplemental source of Pashto lexical information. About 2000 flashcard entries have been collected and added to the Pashto lexicon.

Lexicon Processing and Correction: Existing Pashto and Urdu lexicon resources were improved by spelling normalization, correction of entries with split or run-together words, and the definition of entries for new words.

Pashto: The LDC Pashto lexicon was examined for spelling normalization issues (see Section 0) and general accuracy. Two types of errors were identified: misspelling in the English glosses, and incomplete phrasal translations. When the English definition is a phrase such as "to VERB" or "was VERBing", the Pashto word is often defined with just the word *to* or *was*. There are 90 words glossed as *to* and 28 words glossed as *was*; presumably, many of these represent defective glosses. Additionally, several Pashto words in the lexicon include an attached د /d/ which means *of*; these are sometimes glossed appropriately as "of NOUN", but often they are glossed as just *of*, as shown here:

Pashto String	**Transliteration**	**English Gloss**	**Literal Meaning**	**Correct**
یادو	yadw	remember	remember	yes
یادول	yadwl	to	remember-infinitive	no
دکابل	dkabl	Kabul's	of-Kabul	yes
دامریکا	damryka	of	of-America	no

*Urdu**:*** Work continued on improvements to the LDC Urdu Lexicon. (See Section 0).

Reviewed Lexicon Formats: A review was made of alternatives to the LDC lexicon format, such as the Open Lexicon Interchange Format (OLIF), in order to consider how to integrate additional lexical resources.

2.5.3. Treebanks

Existing French treebank resources were identified, with the most promising being the Abeillé treebank [20] built from 1 million words from the French newspaper *le Monde*.[23]

2.5.4. MT Resources

This section reports the identification of existing language processing software (such as part of speech taggers and morphological analyzers) for Pashto, Farsi, and the African languages.

[23] http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php

Pashto: Existing resources were identified for processing Pashto, including part of speech taggers, morphological analyzers, and rule-based translation systems. (See Appendix A)

***Farsi**:* _Existing resources were identified for the morphological processing of Farsi, for possible use with Dari, a closely related language. (See Appendix B)

African Languages: The African Language Technology (AFLAT) organization was identified as a source of technical papers and other resources for the MT of African languages. (See Appendix C)

Somali was noted to have extensive morphological inflection, which indicates the need to develop a stemming program for this language to improve MT.

2.5.5. Translation

This section reports the development of parallel text resources through the use of human translators for a Pashto/French/English newswire dataset.

***Pashto > French > English**:* The SCREAM lab participated in a project in which French researchers transcribe Pashto news audio (e.g., from Voice of America broadcasts), translating the Pashto text into French text, which is then translated into English in the SCREAM lab. This will eventually create a trilingual corpus of parallel text in Pashto, French, and English.

Initial work was done on one hour of Pashto audio that had been translated into French, and on a separate French sample with a known English reference. Two human translators created the English translations. The metrics multi-bleu and sclite were used to measure variation between the translation and the original reference file, both for quality control and as an assessment of the potential usefulness of the translations as sources for paraphrasing.

English > French: An English file was translated into French to create an additional reference file, considering the possibility that a non-native French speaker might be able to create a translation that contains useful variation from the reference file. The metrics multi-bleu and sclite were used to measure variation between the translation and the original reference.

2.5.6. Summary

Online resources for parallel text were identified for Dari, Pashto, Farsi, and a variety of African languages. French data with hand-edited word alignments were also collected. Existing Pashto and Urdu lexicon files were examined for systematic errors and additional lexical entries were developed. Also, human translators were used to create the English component of a preliminary Pashto/French/English dataset.

The GLOSS website was scraped for independent study courses in 6 languages (Arabic, Dari, Hindi, Pashto, Hausa, and Urdu) and code was created that parsed the files for parallel text. After collection, the online resources were analyzed for problems in language identification and for text extraction problems in PDF files.

2.5.7. Recommendations for Future Work

Language identification should be verified for data collected from online sources. Resources developed from audio news broadcasts should be annotated for voice-overs in different languages.

Parallel text resources collected from online sources should be sorted according to domain (political commentary/newswire vs. immigrant information, for example). News sources should be revisited for additional postings over time. This news data and the continued scraping of the GLOSS website for newer courses and more languages could eventually result in enough parallel corpora to help in training LMs for MT.

2.6 Named Entity Detection Benefits for Language Processing

Section 0 discusses subject matter expertise in computational linguistics and section 0 covers algorithm development for domain specific translation systems.

2.6.1. Computational Linguistics Subject Matter Expertise

The researcher working in this area is familiar with various NE Tagging programs and has experience with different ways to apply NE tagging within the MT pipeline, such as tagging the names with the translation as an attribute, pre-translating the names, or automatically transliterating the names.

2.6.2. Algorithm Development for Domain Specific Translation Systems

Developments for NE processing include the collection of NE lists and methods to correct word boundary errors that impede NE recognition.

Collection of Named Entity Lists: Named Entity lists were collected for Dari and Pashto using election oversight websites, which provide names of persons, media outlets, cities, and provinces in Afghanistan. (See Appendix A)

Word Boundary Correction to Support Named Entity Detection: Arabic script languages often appear with word boundary errors that can impede named entity detection. Some characters are linked, but others are non-joiners, giving the appearance of a space within a word. Sometimes writers fail to type an actual space between words when there are non-joining characters; alternatively, writers may add spaces within words. For example, the Urdu data includes the sequence اوربرازیل /aorbrazyl/ which is the run-together form of اور plus برازیل /aor brazyl/ "and Brazil". Since the letter /r/ does not connect to the following letter, the run-together form preserves the appearance of word spacing.

These variations can cause words to remain unknown in MT in general, and can impede the recognition and translation of named entities in particular. Previous work with the LDC-supplied NE tagger for Urdu revealed problems in the isolation of names from postpositions and punctuation marks. Postpositions were also sometimes tagged as named entities by themselves. (See Appendix D) This shows the importance of applying tokenization and morphological processing to the input files before training or applying a named entity tagger.

The Pashto data show a particularly high frequency of variations involving attached/detached words. For example, the word /d/ ‘of’ is sometimes written separately and sometimes written as part of the following word; similarly, the word /ao/ ‘and’ can be found attached or unattached. Thus, a phrase like "of peace and security" occurs in several variants:

Pashto (Transliterated)	**Meaning**
d soly ao amnyt	of peace and security
d soly aoamnyt	of peace and-security
dsoly ao amnyt	of-peace and security
dsoly aoamnyt	of-peace and-security

Various algorithms were tested to detect and correct these word boundary errors. In previous work, a program was developed to count the frequency of the unigram and bigram variants and convert all forms to the more frequent variant before sending the data to MT. For example, in English the program might note that *snowfall* is more frequent than *snow fall*, and convert all instances to the unigram form.

In the current effort, the bigram-unigram conversion program was applied to Pashto. This process succeeded in splitting instances of preposition and noun, but it also improperly split some characters from stems. Applying the conversion program before MT did not improve the overall performance of the translation system.

Creation of Separate Lexicon for Named Entities: An effort was made to sort the Urdu lexicon into named entities and common nouns, in order to support tagging and translation of named entities. This division was problematic, however, since many names also have common meanings (e.g., *Malik* means *state*; *Aziz* means *beloved*). Future work might resolve these ambiguities by assigning a probability weighting to the various meanings, or by applying LMs to determine contextual meaning.

2.6.3. Summary

Word boundary errors were identified as a primary source of problems for named entity detection and for the automatic creation of named entity lists. Online resources were identified for the manual creation of named entity lists.

2.6.4. Recommendations for Future Work

Future work should compare list-based and statistically-based named entity tagging methods. Morphological issues should be addressed when training and applying named entity detectors in languages with inflected forms or spelling variation.

2.7 Recognition and Translation Performance Assessment

As noted in Section 0, a bigram-unigram conversion program was applied to the Pashto data to split instances of run-together words. This succeeded in isolating previously obscured words and names, however, in some cases characters were improperly split from stems. Applying the conversion program before MT did not improve the overall performance of the translation system.

2.8 English to Urdu Translation

A native-speaker consultant was used to create English/Urdu parallel text resources in two domains, and to review and improve existing word alignments and lexical entries.

2.8.1. Urdu BTEC Dataset

An Urdu version of the BTEC dataset was developed by having a native-speaker consultant translate the English sentences into Urdu. This dataset was then edited and normalized for spelling and punctuation. (See Section 0).

2.8.2. Urdu Information Retrieval Dataset

English sentences and keywords were translated into Urdu by a native-speaker consultant for use in an information retrieval and topic detection project. This dataset was also edited for punctuation normalization. (See Section 0).

2.8.3. Translation Accuracy

This section reports improvements to the Urdu word alignments and the Urdu lexicon, based on the judgments of a human translator.

Word Alignment: The BTEC English/Urdu translations were automatically word-aligned, sorted by alignment score, and returned to the native-speaker consultant for alignment editing of the worst-scoring sentences. This process also helped identify some errors in the initial human translation from English into Urdu, as, for example, when part of a sentence had been omitted in the translation.

Lexicon: Previous work improved the LDC Urdu lexicon via translation of OOV words by a native-speaker consultant. Work continued on this lexicon. The focus was on closed class words (prepositions, conjunctions, numerals) and high frequency words; definitions were created for the most frequent words down to words that occur 4 or more times in our data. (See Appendix D)

A percent unknown program was used to compare the coverage of words in the Urdu training data. Results showed that the augmented dictionary improves coverage over the original LDC dictionary, with the percentage of unknown words dropping from 13% to 5%.

2.8.4. Summary

Additional Urdu resources were created by using a native-speaker consultant to translate documents, edit machine-generated word alignments, and translate individual words for the Urdu lexicon.

2.8.5. Recommendations for Future Work

Automatically-generated word alignment scores proved to be a useful metric for identifying errors in the work of a human translator; this should be added as a routine step in reviewing translation accuracy.

2.8.6. Laboratory System Administration Support

System Administration support was provided to maintain the computational efficacy of the SCREAM Laboratory in order to support SALT research.

Some significant system administration tasks accomplished under this task are listed below. Many frequent system administration tasks, such as routine system maintenance, backups, system repair, system troubleshooting, and user support, also accomplished under this task order may not be listed.

Under the ICER contract, system administration tasks are included in multiple task orders that support the SCREAM Laboratory. In many cases, non-trivial system administration tasks were split between different task orders. Some of the significant tasks listed below may also appear in reports for other task orders as the work was split between multiple task orders.

- Integrated a new 19TB RAID6 arrays into the SCREAM Laboratory network.
- Integrated additional Linux servers and workstations into the SCREAM Laboratory network.
- Updated in-house software to provide near real-time performance monitoring of all the Linux servers and workstations on the SCREAM Lab network. Using ZeroMQ, a high-performance asynchronous messaging library, instant performance statistics from 100 Linux systems are retrieved and displayed within 1-3 seconds.
- Deployed Munin performance monitoring software on all Linux systems. Data collected every 5 minutes and displayed on daily, weekly, and monthly graphs help analyze and identify network and system performance problems.
- Setup virtual tape backups that utilize unused disk space found on the large hard drives in many of the newer Linux workstations, providing better backup coverage for frequently changing data generated by active experiments.
- Analyzed problems with statically linked, legacy, Linux software and large XFS filesystems. Due to the size of the newer RAID 6 arrays in the SCREAM Laboratory, XFS filesystems using 64-bit inodes are being used. Some older binary software that was statically linked with 32-bit system libraries cannot access the 64-bit inodes and will fail with “file not found” errors. In some cases disabling the enable_ino64 parameter on the NFS client will allow the software to access the filesystem.
- Re-configured network connections on the three main SCREAM Lab NFS servers to increase available bandwidth by utilizing multiple Ethernet connections and link aggregation protocols.
- Installed and configured xml-qstat, which provides a web-based user interface to display status of jobs running on Open Grid Scheduler (OGS).
- Updated computer status page to display number and status of OGS jobs. Implemented Javascript pop-up windows to display detailed OGS job information on a per-host basis.
- Moved the SCREAM Lab authentication server from a desktop class system to a rack-mounted server class machine with redundant power supplies.
- Tested Centrify Express, software that provides Linux integration into Microsoft Active Directory, for possible use as a Single Sign-On (SSO) solution and upgrade to NIS and Samba authentication schemes.
- Performed some initial experiments with the Ceph distributed storage system.
- Created a tutorial for generating screencast videos from Linux using FFmpeg or gtk-recordMyDesktop. Determined a method to multiplex audio from a running application and the microphone to provide narration to a screencast video along with audio from the application.

3.0 CONCLUSIONS AND RECOMMENDATIONS

This document summarized work completed by SRA International, Inc. during the period 20 August 2009 to 28 February 2013.

The Sphinx-4 speech recognition engine was modified to apply class-based feature transforms and correctly mark LM scores in lattices. An Arabic ASR system was developed on the TDT4 corpus. Narrowband and wideband AMs were developed using HTK, and a trigram LM was estimated using the SRILM Toolkit. An English ASR system was developed for the IWSLT 2011 evaluation. An improvement in system performance was obtained by discriminatively training the HMMs, applying SAT, decoding with an interpolated LM, and rescoring with a 4-gram LM.

HDecode recognition lattices typically have less hypothesized word sequences than Sphinx-4 lattices. It might be beneficial to investigate alternative algorithms for creating the lattices. The Arabic ASR system did not apply SAT because the speaker identities were not known. One possible workaround would be to automatically cluster the segments. Another potential area for improvement would be using additional text data to estimate the LM. No improvement in WER was obtained on the IWSLT dev2010 partition by combining the outputs from three different ASR systems. This may be because the difference in WER between the best and worst system is 9.9% absolute. An improvement in system performance might be obtained by combining systems with more comparable WERs or using a different method for system combination.

Various improvements were developed for the SMT2 SMT system. Character level processing was improved by reducing variation with spelling normalization and tokenization. Morphological processing was used to improve recognition of different instances of the same word. Errors in sentence alignment were detected and corrected. Word alignment errors were analyzed to reduce phrase table errors. Several paraphrasing software programs were explored to generate variant sentences with similar meanings for use in training or scoring of MT.

Performance enhancements, functionality enhancements, and error corrections were performed on various software packages which are actively being used in the SCREAM Laboratory. The Carnegie Mellon MT Eval server had various scoring displays and functions added to its framework.

Spelling normalization for Dari should be investigated further, since there are variant characters that occur in similar proportions in the data, according to the preference of the writer. Pashto machine translation might be improved by creating a Pashto morphological analyzer, possibly built within the GF/Molto rule system, using published morphological rules for Pashto noun and verb inflection.

Lattice and confusion net representations were explored for paraphrasing; lattices could also be applied to represent other variations, including punctuation variants (e.g., French « » vs. "), inflected forms, and semantic alternatives such as the simple vs. compound past forms of French verbs.

Current punctuation normalization and tokenization methods depend on well-defined lists of punctuation characters. However, data across languages show unexpected use of punctuation, and this variation increases in informal text such as blogs and social media. Future work should therefore consider more dynamic methods of recognizing and normalizing punctuation variation.

Sentence alignment corrections did not provide the expected benefits for translation. Instead, further research should be focus on word alignment and its relationship to phrase table extraction, including the use of word alignment weights and the possible use of validated word alignments to delimit phrases.

Performance enhancements should be continually researched, using both traditional techniques and utilizing a GPGPU.

Three English HMM-based speech synthesis systems were developed using full context labels. A GUI was developed for recording speech and adapting an HMM-based speech synthesis system.

English HMM-based speech synthesis systems were only developed using male speakers; it would be useful to train systems on female speakers as well. In addition, it could be beneficial to conduct a formal listening test to compare the systems.

The GUI used to record speech currently relies on the Snack Sound Toolkit. The most recent release of this toolkit from the Royal Institute of Technology (KTH) was in 2005, and on our newest computers we have experienced problems interfacing with the sound card. It would be worthwhile to implement an alternative method for recoding and playing audio.

Online resources for parallel text were identified for Dari, Pashto, Farsi, and a variety of African languages. The GLOSS website was scraped for independent study courses in six languages (Arabic, Dari, Hindi, Pashto, Hausa, and Urdu) and code was created that parsed the files for parallel text.

Parallel text resources collected from online sources should be sorted according to domain (political commentary/newswire vs. immigrant information, for example). News sources should be revisited for additional postings over time. This news data and the continued scraping of the GLOSS website for newer courses and more languages could eventually result in enough parallel corpora to help in training LMs for MT.

Word boundary errors were identified as a primary source of problems for named entity detection and for the automatic creation of named entity lists. Online resources were identified for the manual creation of named entity lists.

Future work should compare list-based and statistically-based named entity tagging methods. Morphological issues should be addressed when training and applying named entity detectors in languages with inflected forms or spelling variation.

Additional Urdu resources were created by using a native-speaker consultant to translate documents, edit machine-generated word alignments, and translate individual words for the Urdu lexicon.

Lastly, automatically-generated word alignment scores proved to be a useful metric for identifying errors in the work of a human translator; this should be added as a routine step in reviewing translation accuracy.

4.0 REFERENCES

1. W. Walker *et al.*, “Sphinx-4: A Flexible Open Source Framework for Speech Recognition,” Sun Microsystems Inc., SMLI TR2004-0811, 2004.
2. Cambridge University Engineering Department, “The HTK Book,” 2009 (Available at http://htk.eng.cam.ac.uk).
3. A. Stolcke, “SRILM – An Extensible Language Modeling Toolkit,” in *Proc. of the International Conference on Spoken Language Processing,* Denver, Colorado, 2002.
4. “1997 Mandarin Broadcast News Speech (HUB4-NE),” *Linguistic Data Consortium*, Philadelphia, 1998 (Available at http://www.ldc.upenn.edu).
5. J. Kong and D. Graff, “TDT4 Multilingual Broadcast News Speech Corpus,” *Linguistic Data Consortium*, Philadelphia, 2005 (Available at http://www.ldc.upenn.edu).
6. D. Graff *et al.*, “1996 English Broadcast News Speech (HUB4),” *Linguistic Data Consortium*, Philadelphia, 1997 (Available at http://www.ldc.upenn.edu).
7. J. Fiscus *et al.*, “1997 English Broadcast News Speech (HUB4),” *Linguistic Data Consortium*, Philadelphia, 1998 (Available at http://www.ldc.upenn.edu).
8. L. Lamel, “Lightly Supervised and Unsupervised Acoustic Model Training,” *Computer Speech and Language*, **Vol.** 16, pp. 115–129, 2002.
9. D. Povey *et al.*, “The Kaldi Speech Recognition Toolkit,” *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Hawaii, USA.
10. M. Humayoun, et al., "Urdu Morphology, Orthography and Lexicon Extraction." *CAASL-2: The Second Workshop on Computational Approaches to Arabic Script-based Languages*, July 21-22, 2007, LSA 2007 Linguistic Institute, Stanford University, 2007.
11. F. Zuhra and M. Khan, "A Corpus-Based Finite State Morphological Analyzer for Pashto" in *Proceedings of the Conference on Language & Technology*, 2009.
12. C. Bannard and C. Callison-Burch, "Paraphrasing with Bilingual Parallel Corpora", *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 2005.
13. M. Denkowski and A. Lavie, "METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support For Five Target Languages", *Proceedings of the ACL 2010 Joint Workshop on Statistical Machine Translation and Metrics MATR*, 2010.
14. J. Ganitkevitch et al., "Joshua 4.0: Packing, PRO, and Paraphrases", *Proceedings of the 7th Workshop on Statistical Machine Translation*, Montréal, Canada, June 7-8, 2012.
15. Philipp Koehn and Jean Senellart, “Convergence of Translation Memory and Statistical Machine Translation”, *AMTA Workshop on MT Research and the Translation Industry,* Denver Colorado, Oct 31 – Nov 4, 2010.
16. J. Garofolo *et al.*, “CSR-I (WSJ0) Complete,” *Linguistic Data Consortium*, Philadelphia, 2007 (Available at http://www.ldc.upenn.edu).
17. “CSR-II (WSJ1) Complete,” *Linguistic Data Consortium*, Philadelphia, 1994 (Available at http://www.ldc.upenn.edu).
18. F. Och and H. Ney, "Improved Statistical Alignment Models". *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hongkong, China, October 2000.

19. J. Graça et al., "Building a golden collection of parallel Multi-Language Word Alignment", *The 6th International Conference on Language Resources and Evaluation* (LREC 2008), May 2008.

20. Abeillé , A., et al., "Building a Treebank for French" in Abeillé , A., ed., *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer, 2003.

APPENDIX A - Dari and Pashto Parallel Text Resources

Parallel Text and Other Resources: English/Dari/Pashto

DRAFT: revised 21 December 2012

Introduction

This documents is divided into sections for:

Parallel Text
- News
- Documents > 20 pages
- Websites with parallel pages, named in parallel
- Lesser Resources

Dictionaries and Wordlists
Named Entities
News: Comparable Text
Multimedia
Tools for MT
Sites to Watch

Within these sections, there are some sub-sections for resources that are two-way parallel only (e.g., just English/Dari). These are highlighted in magenta.

General notes for working with Dari and Pashto websites

Green highlight indicates instructions for the data collector. Yellow highlight indicates domain such as news, medical, etc. (data should be saved in separate files for each domain).

Navigation
Abbreviations for Dari include dr and fa (for Farsi). Abbreviations for Pashto include pa and ps.
On websites with parallel text, there is often a tab at the top or bottom for selecting the language with these words:

English دری پښتو
[English Dari Pashto]

Distinguishing Dari and Pashto
Characters with rings (ټ ڼ ړ ګ) are Pashto (but might show up in Dari text as part of named entities).
Dari tends to use digits like 123, while Pashto tends to use digits like ١ ٢ ٣

Dates
Dari and Pashto documents may express dates in the Islamic calendar (2010 = 1388, 1389)

Language Names
The language of Iran was called Persian, and is now usually called Farsi. Dari is the version of Persian/Farsi that is spoken in Afghanistan.

Pashto may also be listed as Pushto, Pukhto, and other variations (see Kopris, "Computing in Pashto" 2005 http://www.iranianlinguistics.org/papers/Kopris_Proof_2.22.pdf)

Parallel Text

News

Sada-e-Azadi newspaper
parallel text in English, Dari, Pashto, in columns
www.sada-e-azadi.net

Institute for War & Peace Reporting
http://iwpr.net/
There appear to be about 100 news articles in English, Dari, and Pashto, with ongoing postings.
"Stories written by trainee journalists are published in the Afghan Recovery Report, which appears each week on IWPR's website in English, Dari and Pashto ..."
http://iwpr.net/programme/142/issue-archive [most recent articles are only in English]
http://iwpr.net/programme/142/issue-archive?page=2 [starting in April 2012, some articles translated into Dari and Pashto]
Articles are not named in parallel, but there are links at the top of each article to the alternate language versions.
To get just the Pashto articles:
http://iwpr.net/ps/programme/%D8%A7%D9%81%D8%BA%D8%A7%D9%86%D8%B3%D8%AA%D8%A7%D9%86-afghanistan-pashto

Pajhwok newspaper -- registration required --
parallel articles in English, Dari, Pashto
http://www.pajhwok.com/en
http://www.pajhwok.com/dr
http://www.pajhwok.com/ps
-- but some older articles are freely available --
for example, on 7 June 2012, articles prior to 25 May 2012 are available
Search by date here: http://www.pajhwok.com/en/archives
from the English article, click on the link(s) in the upper right for Dari دری or Pashto پښتو version
http://www.pajhwok.com/en/2012/05/24/uruzgan-officials-donate-blood
not all English articles have translations; not named in parallel

USAID press releases, many translated into English/Dari/Pashto
http://afghanistan.usaid.gov/en/newsroom/press_releases
Starting in May of 2011, just English titles are listed, but at the bottom of the English article are links to pdf files for English, Dari, and Pashto, e.g.,
2011.10.04 Press Release Women's Capacity Building Graduation (English)
2011.10.04 Press Release Women's Capacity Building Graduation (Pashto)

2011.10.04 PressRelease-Women's Capacity Building Graduation (Dari)

MRRD – Afghanistan Ministry of Rural Rehabilitation and Development
press releases in English/Dari/Pashto since May 5, 2010
note: English is unidiomatic
http://mrrd.gov.af/en/news
for example,
http://mrrd.gov.af/en/news/3772
http://mrrd.gov.af/fa/news/3772
http://mrrd.gov.af/ps/news/3772

MCIT - Afghanistan Ministry of Communications and Information Technology
press releases in English/Dari/Pashto since Dec 2010
http://mcit.gov.af/en/news
for example,
http://mcit.gov.af/en/news/3485
http://mcit.gov.af/fa/news/3485
http://mcit.gov.af/ps/news/3485

World Bank press releases
62 parallel articles, but not named in parallel.
Index pages for Dari and Pashto:
http://web.worldbank.org/WBSITE/EXTERNAL/NEWS/0,,lang:434533~menuPK:34467~pagePK:64254793~piPK:64254755~theSitePK:4607,00.html
http://web.worldbank.org/WBSITE/EXTERNAL/NEWS/0,,lang:434614~menuPK:34467~pagePK:64254793~piPK:64254755~theSitePK:4607,00.html
for example,
http://web.worldbank.org/WBSITE/EXTERNAL/NEWS/0,,contentMDK:22988862~menuPK:34464~pagePK:34370~piPK:34424~theSitePK:4607,00.html
http://siteresources.worldbank.org/NEWS/PressRelease/22988818/FinancialSRRPDari.pdf
http://siteresources.worldbank.org/NEWS/PressRelease/22988870/FinancialSRRPPashto.pdf

Documents > 20 pages

** news domain **
Asia Foundation
Afghanistan in 2007 (200 pp)
www.asiafoundation.org/publications/pdf/20 [English]
www.asiafoundation.org/publications/pdf/255 [Dari]
www.asiafoundation.org/publications/pdf/256 [Pashto]

** medical domain **
Canada, Community Resource Connections of Toronto

Navigating Mental Health Services in Toronto - A Guide for Newcomer Communities
(100 page brochure)
http://www.crct.org/lanresources/PDFs/CRCT-English-041511-web.pdf
http://www.crct.org/lanresources/PDFs/CRCT-NMHS-Dari.pdf
http://www.crct.org/lanresources/PDFs/CRCT-NMHS-Pashto.pdf

** citizen information domain **
First Steps: An Orientation Package for Government-Assisted Refugees
(103 page brochure)
http://www.settlement.org/downloads/First_Steps_English.pdf
http://www.settlement.org/downloads/First_Steps_Dari.pdf
http://www.settlement.org/downloads/First_Steps_Pashtu.pdf

** legal domain **
UNIFEM – UN Development Fund for Women
Parliamentary Manual (80 pp)
http://www.unifem.org/afghanistan/docs/pubs/05/parliament_manual_EN_05.pdf
http://www.unifem.org/afghanistan/docs/pubs/05/parliament_manual_DR_05.pdf
http://www.unifem.org/afghanistan/docs/pubs/05/parliament_manual_PSH_05.pdf

** news domain **
Mine Action Coordination Centre of Afghanistan, reports
http://www.macca.org.af/file.php?id=408 [English, annual report, 1389] 59 pp.
http://www.macca.org.af/file.php?id=414 [Dari, annual report, 1389]
http://www.macca.org.af/file.php?id=413 [Pashto, annual report, 1389]

http://www.macca.org.af/file.php?id=308 [English, Integrated Operational Framework, 1390] 67 pp.
http://www.macca.org.af/file.php?id=311 [Dari, Integrated Operational Framework, 1390]
http://www.macca.org.af/file.php?id=310 [Pashto, Integrated Operational Framework, 1390]

** political domain **
NDI – National Democratic Institute
observation of 2010 election (21 pp)
www.ndi.org/files/NDI_Afghan_2010_EOM_preliminary_statement.pdf
www.ndi.org/files/NDI_Afghan_2010_EOM_preliminary_statement-Dari.pdf
www.ndi.org/files/NDI_Afghan_2010_EOM_preliminary_statement-Pashto.pdf
similarly, this document (14 pp)
www.ndi.org/files/Afghanistan_EOM_Preliminary_Statement.pdf
www.ndi.org/files/NDI_Afghanistan_EOM_Preliminary_statement%20_Dari.pdf
www.ndi.org/files/NDI_Afghanistan_EOM_Statement%20_Pashto.pdf
polling manual (26 pages, includes charts, questionnaires)
www.ndi.org/files/Afghanistan_Polling_Agent_Manual2010.pdf
www.ndi.org/files/Afghanistan_Polling_Agent_Manual2010-Dari.pdf
www.ndi.org/files/Afghanistan_Polling_Agent_Manual2010-Pashto.pdf

** political domain **
IEC – Independent Election Commission
www.iec.org.af/jemb.org/index.html
Media Commission Report (25 pp)
[note: Dari document has tracked changes, might show up in extracted text]
www.iec.org.af/jemb.org/pdf/JEMBS_LGL_MC_Final_Report_2005-11-18_Eng.pdf
www.iec.org.af/jemb.org/pdf/JEMBS_LGL_MC_Final_Report_2005-11-18_Dari.pdf
www.iec.org.af/jemb.org/pdf/JEMBS_LGL_MC_Final_Report_2005-11-18_Pas.pdf

** legal domain **
Electoral Law (25 pp)
[note: Dari and Pashto are in columns in the second document, Dari on the left]
www.iec.org.af/pdf/legalframework/law/electorallaw_eng.pdf
www.iec.org.af/pdf/legalframework/law/electorallaw.pdf

Documents > 20 pages, English/Dari only

** legal domain **
UNIFEM – UN Development Fund for Women
documents are indexed by year; documents prior to 2006 are scanned only
http://www.unifem.org/afghanistan/media/pubs/year.php?pubYear=2007
http://www.unifem.org/afghanistan/media/pubs/year.php?pubYear=2006
Toolkit 1: International Instruments for the Protection of Women's Human Rights, 40 pp [2007]
Toolkit 2: Women Parliamentarians Making a Difference in Politics, 44 pp [2007]
Toolkit 3: Legislative Process, 26 pp [2007]
Uncounted and Discounted: A Secondary Research Project on Violence Against Women in Afghanistan, 53 pp [2006]

** medical domain **
Canada, Center for Addiction and Mental Health
http://www.camh.net/About_Addiction_Mental_Health/Mental_Health_Information/Alone_in_Canada/alone_in_canada.pdf
http://www.camh.net/About_Addiction_Mental_Health/Mental_Health_Information/Alone_in_Canada/alone_in_canada_dari.pdf [70 page brochure]

Documents > 20 pages, English/Pashto only

** news domain **
Mine Action Coordination Centre of Afghanistan, reports
http://www.macca.org.af/file.php?id=191 [English, annual report, 1388] 55 pp.
http://www.macca.org.af/file.php?id=193 [Pashto, annual report, 1388]

** medical domain **

Centre for Addiction and Mental Health
http://www.problemgambling.ca/EN/Documents/Guide%20for%20Families%202006.pdf
http://www.problemgambling.ca/EN/Documents/2844_GuideForFamilies_PASHTO.pdf
[48 page brochure]

Websites with parallel pages, named in parallel

** political domain **
IEC – Independent Election Commission
[note: some of the pdfs are text, but some are only scanned documents]
parallel pages, for example:
www.iec.org.af/eng/content.php?id=1&cnid=4
www.iec.org.af/dari/content.php?id=1&cnid=4
www.iec.org.af/pashto/content.php?id=1&cnid=4
2010 press releases, index
www.iec.org.af/eng/content.php?id=6&cnid=28
www.iec.org.af/dari/content.php?id=6&cnid=28
www.iec.org.af/pashto/content.php?id=6&cnid=28
for example,
www.iec.org.af/pdf/wolesi-pressr/pr_statement_on_staff_turnout_and_materials.pdf
www.iec.org.af/pdf/wolesi-pressr/dari/pr_statement_on_staff_turnout_and_materials.pdf
www.iec.org.af/pdf/wolesi-pressr/pashto/pr_statement_on_staff_turnout_and_materials.pdf
2009 press releases, index
http://www.iec.org.af/eng/content.php?id=6&cnid=27
http://www.iec.org.af/dari/content.php?id=6&cnid=27
http://www.iec.org.af/pashto/content.php?id=6&cnid=27
for example,
http://www.iec.org.af/pdf/pressrelease/preperations_runoff_20091029.pdf
http://www.iec.org.af/pdf/pressrelease_dari/preperations_runoff_20091029Dari.pdf
http://www.iec.org.af/pdf/pressrelease_dari/preperations_runoff_20091029Pashto.pdf

** political domain **
IEC – Independent Election Commission, JEMB – Joint Electoral Management Body
note: page names here are not always exactly parallel.
best method: select an English page, then click on the English/Dari/Pashto buttons.
www.iec.org.af/jemb.org/eng/jembbg.html
www.iec.org.af/jemb.org/pashto/jembbg.html
www.iec.org.af/jemb.org/dari/jembbg.html
factsheets
www.iec.org.af/jemb.org/eng/bg&factsheets.html
www.iec.org.af/jemb.org/dari/bg&factsheets.html
www.iec.org.af/jemb.org/pashto/bg&factsheets.html
FAQ -–useful for examples of question structures--

www.iec.org.af/jemb.org/eng/parliamentary_faq.html
www.iec.org.af/jemb.org/pashto/faqs.html
www.iec.org.af/jemb.org/dari/faqs.html

** political domain **
IEC -JEMB – Joint Electoral Management Body – Media Commission
www.iec.org.af/jemb.org/media_commission/background.html
www.iec.org.af/jemb.org/media_commission/dari/background.html
www.iec.org.af/jemb.org/media_commission/pashto/background.html
(see also Named Entity section, below)

** news domain **
UNDP – United Nations Development Programme
20 newsletters, approx. 10 pp. each
www.undp.org.af/Publications/Newsletters/english/UNDP%20Newsletter-%20MarchApril08-English.pdf
www.undp.org.af/Publications/Newsletters/dari/UNDP%20Newsletter-%20MarchApril08-Dari.pdf
www.undp.org.af/Publications/Newsletters/pashto/UNDP%20Newsletter-%20MarchApril08-Pashto.pdf
various pages in parallel, e.g., 9 development goals
www.undp.org.af/MDGs/goal1.htm
www.undp.org.af/Dari/d_MDGs/d_goal1.htm
www.undp.org.af/Pashto/p_MDGs/p_goal1.htm

** political domain **
NDI- National Democratic Institute
election observations – 14 newsletters, 2pp. each
www.ndi.org/node/16535 [index]
for example,
www.ndi.org/files/Afghanistan_Elections_Update_Issue1.pdf
www.ndi.org/files/Afghanistan_Elections_Update_Issue1-Dari.pdf
www.ndi.org/files/Afghanistan_Elections_Update_Issue1-Pashto.pdf

** what domain ? **
ATRA – Afghanistan Telecom Regulatory Authority
www.atra.gov.af
www.atra.gov.af/dr/home.html
www.atra.gov.af/pa/home.html
parallel pages – some are just English/Pashto, some are just English/Dari
www.atra.gov.af/en/History.html
www.atra.gov.af/pa/History.html
www.atra.gov.af/en/3rd_rtd_uas_projects.html
www.atra.gov.af/pa/3rd_rtd_uas_projects.html

Websites with parallel pages, named in parallel - English/Dari only

** general interest domain **
http://langmedia.fivecolleges.edu/culturetalk/afghanistan/index.html
Five College Center for the Study of World Languages
Dari and English transcripts of video interviews on 45 cultural topics
(see multimedia entry, below)

Lesser Resources: Websites with only a few parallel pages, or parallel pages that are not named systematically

ICTJ - International Center for Transitional Justice
a series of 2-page fact sheets
www.ictj.org
Conflict and Transitional Justice in Africa
http://ictj.org/sites/default/files/ICTJ-Africa-Conflict-Facts-2009-English.pdf
http://ictj.org/sites/default/files/ICTJ-Africa-Conflict-Facts-2009-Dari.pdf
http://ictj.org/sites/default/files/ICTJ-Africa-Conflict-Facts-2009-Pashto.pdf
Truth and Reconciliation in Morocco
http://ictj.org/sites/default/files/ICTJ-Morocco-TRC-2009-English.pdf
http://ictj.org/sites/default/files/ICTJ-Morocco-TRC-2009-Dari.pdf
http://ictj.org/sites/default/files/ICTJ-Morocco-TRC-2009-Pashto.pdf
Pursuing Peace, Justice, or Both?
http://ictj.org/sites/default/files/ICTJ-Global-Peace-Justice-2009-English.pdf
http://ictj.org/sites/default/files/ICTJ-Global-Peace-Justice-2009-Dari.pdf
http://ictj.org/sites/default/files/ICTJ-Global-Peace-Justice-2009-Pashto.pdf
A Transition in Nepal from Insurgency to Governing
http://ictj.org/sites/default/files/ICTJ-Nepal-Insurgency-Governing-2009-English.pdf
http://ictj.org/sites/default/files/ICTJ-Nepal-Insurgency-Governing-2009-Dari.pdf
http://ictj.org/sites/default/files/ICTJ-Nepal-Insurgency-Governing-2009-Pashto.pdf
Transitional Justice in the Former Yugoslavia
http://ictj.org/sites/default/files/ICTJ-FormerYugoslavia-Justice-Facts-2009-Dari.pdf
http://ictj.org/sites/default/files/ICTJ-FormerYugoslavia-Justice-Facts-2009-Dari.pdf
http://ictj.org/sites/default/files/ICTJ-FormerYugoslavia-Justice-Facts-2009-Pashto.pdf
An Overview of Conflict in Columbia
http://ictj.org/sites/default/files/ICTJ-Colombia-Conflict-Facts-2009-English.pdf
http://ictj.org/sites/default/files/ICTJ-Colombia-Conflict-Facts-2009-Dari.pdf
http://ictj.org/sites/default/files/ICTJ-Colombia-Conflict-Facts-2009-Pashto.pdf
What Next for International Justice?
http://ictj.org/sites/default/files/ICTJ-Global-ICC-Prosecutions-2009-English.pdf
http://ictj.org/sites/default/files/ICTJ-AW-Afghanistan-Prosecutions-2009-Dari.pdf
http://ictj.org/sites/default/files/ICTJ-AW-Afghanistan-Prosecutions-2009-Pashto.pdf

USAID – US Agency for International Development
8 brochures, 2 pp each
http://afghanistan.usaid.gov/en/Programs.aspx

for example,
http://afghanistan.usaid.gov/proxy/Document.790.aspx
http://afghanistan.usaid.gov/proxy/Document.847.aspx [Dari]
http://afghanistan.usaid.gov/proxy/Document.848.aspx [Pashto]
Index to press releases, not named in parallel prior to May 2011
http://afghanistan.usaid.gov/en/newsroom/press_releases
for example, seed distribution
http://afghanistan.usaid.gov/en/USAID/Article/1258/Afghan_Farmers_Break_Ground_Sow_Seeds_in_AVIPAs_2010_Wheat_Seed_Distribution_Launch
http://afghanistan.usaid.gov/en/USAID/Article/1259 [Dari]
http://afghanistan.usaid.gov/en/USAID/Article/1260 [Pashto]

ECC –Election Complaints Commission
FAQ – useful for examples of question structures
www.ecc.org.af/en/index.php?option=com_content&view=article&id=53&Itemid=68
www.ecc.org.af/dari/index.php?option=com_content&view=article&id=74&Itemid=73
www.ecc.org.af/pashtu/index.php?option=com_content&view=article&id=66&Itemid=76

NSP – National Solidarity Program
www.nspafghanistan.org
www.nspafghanistan.org/indexdari.aspx
www.nspafghanistan.org/indexpashto.aspx
19 page report
www.nspafghanistan.org/files/2nd%20QR%201388%20-Main%20Text.doc
www.nspafghanistan.org/files/گزارش%20ربع%20دوم%20متن%20اصلی.doc [Dari]
www.nspafghanistan.org/files/د%20ملي%20پيوستون%20پروگرام.doc [Pashto]
FAQ – useful for question structures
www.nspafghanistan.org/default.aspx?Sel=26
note: all three versions of the FAQ have the same address; on this site you have to click on the language tabs to put pages into English, Dari, or Pashto

US Department of State: Secretary Clinton's message to the people of Pakistan
http://www.state.gov/secretary/rm/2009a/12/133087.htm
http://www.state.gov/documents/organization/133299.pdf
http://www.state.gov/documents/organization/133308.pdf

US White House: speeches from whitehouse.gov
Joint Statement from the President and President Karzai of Afghanistan
http://www.whitehouse.gov/the-press-office/joint-statement-president-and-president-karzai-afghanistan
http://www.whitehouse.gov/sites/default/files/rss_viewer/joint_statement_dari.pdf
http://www.whitehouse.gov/sites/default/files/rss_viewer/joint_statement_pashto.pdf
Remarks by the President in Address to the Nation on the Way Forward in Afghanistan and Pakistan
http://www.whitehouse.gov/issues/defense/afghanistan

http://www.whitehouse.gov/sites/default/files/091201-obama-afghanistan-speech-dari.pdf
http://www.whitehouse.gov/sites/default/files/091201-obama-afghanistan-speech-pashto.pdf
Fact Sheet: The Way Forward in Afghanistan
http://www.whitehouse.gov/the-press-office/way-forward-afghanistan
http://www.whitehouse.gov/sites/default/files/FACT_SHEET_DARI.pdf
http://www.whitehouse.gov/sites/default/files/FACT_SHEET_PASHTO.pdf
President Obama in Jakarta: "Indonesia's Example To the World"
http://www.whitehouse.gov/the-press-office/2010/11/10/remarks-president-university-indonesia-jakarta-indonesia
http://www.whitehouse.gov/files/documents/20101110-potus-university-indonesia-dari.pdf
http://www.whitehouse.gov/files/documents/20101110-potus-university-indonesia-pashto.pdf
REMARKS BY THE PRESIDENT ON A NEW BEGINNING (Cairo)
http://www.whitehouse.gov/the-press-office/remarks-president-cairo-university-6-04-09
http://www.whitehouse.gov/files/documents/anewbeginning/SPEECH_as_delivered-Dari.pdf
http://www.whitehouse.gov/files/documents/anewbeginning/SPEECH_as_delivered-Pashto.pdf
Statement by the President on the Occasion of Ramadan [several documents like this]
http://www.whitehouse.gov/the-press-office/2010/08/11/statement-president-occasion-ramadan
http://www.whitehouse.gov/sites/default/files/rss_viewer/Ramadan_Dari.pdf
http://www.whitehouse.gov/sites/default/files/rss_viewer/Ramadan_Pashto.pdf
additional translated remarks can be found by searching on the language name:
http://www.whitehouse.gov/search/site/dari
http://www.whitehouse.gov/search/site/pashto

AIHRC – Afghanistan Independent Human Rights Commission
strategic action plan
www.aihrc.org.af/2010_eng/Eng_pages/About/Action_Plan_E/Strategic_Action_Plans.pdf
www.aihrc.org.af/2010_pashto/Strategic%20and%20Action%20Plans%20in%20Pashto.pdf
www.aihrc.org.af/2010_dari/Dri_Pages/Action_Plan/Action_Plan_Dri.pdf

Mine Action Programme of Afghanistan – MACCA – Mine Action Coordination Centre
monthly newsletter, about 5 pp, English/Dari/Pashto, from Nov. 2008 on
www.macca.org.af/en/Newsletter_2010.html
www.macca.org.af/dr/Newsletter_2010.html
www.macca.org.af/pa/Newsletter_2010.html
there are also some press releases, 2008-2011
http://www.macca.org.af/en/press_release_2011.html [English index]

http://asiafoundation.org/publications/index.php?q=&searchType=country&country=1

various short documents in addition to the large one listed in the documents section, above

National Olympic Committee of the Islamic Republic of Afghanistan
a handful of pages about the Olympics
www.nocafghanistan.com
www.nocafghanistan.com/athens2004_en.html
www.nocafghanistan.com/athens2004_fa.html
www.nocafghanistan.com/athens2004_ps.html

a Pajhwok election site, not part of the main newspaper (subscription) site:
www.pajhwokelections.af
www.pajhwokelections.af/pashto
www.pajhwokelections.af/dari
main news section articles are not parallel, but articles on column menu are parallel:
www.pajhwokelections.af/inaugural_speech.php
www.pajhwokelections.af/dari/inaugural_speech.php
www.pajhwokelections.af/pashto/inaugural_speech.php

Supreme Court of Afghanistan
has many files listed with links for English, Dari, Pashto, but often these just lead to a Dari version
http://www.supremecourt.gov.af/en/documents
Laws of the Organization and Authority…
http://supremecourt.gov.af/Content/Media/Documents/Law_on_Org_juris_courts_English112011121448474.pdf [English]
http://supremecourt.gov.af/Content/Media/Documents/OG_0851_org_juris112011122525641.pdf [Dari/Pashto in columns]

UNIFEM – UN Development Fund for Women
3 short documents
www.unifem.org/afghanistan/media/pubs/year.php?pubYear=2005

HRIS – Health Rights Information Scotland – 2 brochures
http://www.hris.org.uk/patient-information/other-languages-and-formats/translations/pashto/
www.hris.org.uk/mod_product/FileUpload/Uploads/742_entitlements_Pashto.jpg
[English/Pashto]
http://www.hris.org.uk/resources/asylum-pashto/ [English/Pashto] (also audio, see multimedia section)

Afghan Parliament
http://wj.parliament.af/english.aspx
http://wj.parliament.af/Default.aspx
http://wj.parliament.af/pashto.aspx
short news articles, ongoing, since Feb. 2008

http://wj.parliament.af/pve/showdoc.aspx?Id=1004
http://wj.parliament.af/pvd/showdoc.aspx?Id=2279
but no way to match up the articles, except matching photos

UNAMA –UN Assistance Mission in Afghanistan
some parallel press releases scattered throughout website
http://unama.unmissions.org/Default.aspx?tabid=1762
http://unama.unmissions.org/Default.aspx?tabid=1760
for example, Peace Day
http://unama.unmissions.org/Default.aspx?tabid=1760&ctl=Details&mid=2002&ItemID=5782
http://unama.unmissions.org/Portals/UNAMA/Press%20Releases/UNAMA%20PRESS%20RELEASE%
http://unama.unmissions.org/Portals/UNAMA/Press%20Releases/UNAMA%20PRESS%20RELEASE%

Afghanistan Legal Documents Exchange Center
http://afghantranslation.checchiconsulting.com
some documents in both English and Dari, a few also in Pashto. some are just scanned text.

Ministry of Commerce, some parallel pages, hard to locate
http://commerce.gov.af/english/pdf%20files/RFP-Wheat_transportation-India-Afg_English.pdf http://commerce.gov.af/Dari/Dari/pdf%20files/RFP-Wheat_Transportation-India-Afg_Dari.pdf

Afghanistan Civil Society Forum
website has English, Dari, and Pashto sections, but it is not clear if the same things exist in each section.
www.acsf.af/English/reports/2009AnnualReport.pdf [English version, 96 pp]
not sure if there are Dari and Pashto versions here somewhere

Afghanistan Centre at Kabul University
www.acku.edu.af/?lang=english
www.acku.edu.af/?lang=da
www.acku.edu.af/?lang=pa
hard to match articles; Pashto not always available
example:
www.acku.edu.af/?p=news&nid=55 [English – September]
www.acku.edu.af/?lang=da&p=news&nid=53 [Dari – September]
example:
www.acku.edu.af/?p=acku_building [English]
www.acku.edu.af/?lang=da&p=ساختمان%20نو [Dari]

Rights and Democracy
http://rightsanddemocracyaf.org/index.htm
a few parallel items

Human Rights Watch
www.hrw.org/en/languages?filter0=gbz
not clear if articles are parallel

Ministry of Education
http://english.moe.gov.af
no easy way to match articles

The Funders' Network for Afghan Women (FNAW)
www.funders-afghan-women.org/
www.funders-afghan-women.org/index.php?lid=2
www.funders-afghan-women.org/index.php?lid=3
some parallel pages in English and Dari
www.funders-afghan-women.org/index.php?page=supporting [English and Dari – use tabs]

Interniche: against animal experimentation
http://www.interniche.org/index1.html
http://www.interniche.org/intro/languages/pashto.html
http://www.interniche.org/intro/languages/dari.html

Office of the Immigration Services Commissioner
http://oisc.homeoffice.gov.uk/about_oisc/publication_scheme/oisc_documents/
Complaints Form
http://oisc.homeoffice.gov.uk/assets/uploads/publications/OISC_English.pdf
http://oisc.homeoffice.gov.uk/assets/uploads/publications/OISC_Dari.pdf
http://oisc.homeoffice.gov.uk/assets/uploads/publications/OISC_Pashto.pdf
General Info
http://oisc.homeoffice.gov.uk/assets/uploads/publications/English%20-%20Revised%20General%20Info%202009.pdf
http://oisc.homeoffice.gov.uk/assets/uploads/publications/English%20-%20Revised%20General%20Info%202009.pdf
http://oisc.homeoffice.gov.uk/assets/uploads/publications/OISC%20factsheet_Pashto_Web.pdf

health translations - Victoria, Australia
privacy
http://www.health.vic.gov.au/pcps/downloads/coordination/translations/privacy_english.pdf
http://www.health.vic.gov.au/pcps/downloads/coordination/translations/privacy_dari.pdf
http://www.health.vic.gov.au/pcps/downloads/coordination/translations/privacy_pushto.pdf
consent

http://www.health.vic.gov.au/pcps/downloads/coordination/translations/consent_english.pdf
http://www.health.vic.gov.au/pcps/downloads/coordination/translations/consent_dari.pdf
http://www.health.vic.gov.au/pcps/downloads/coordination/translations/consent_pushto.pdf

Lesser Resources - English/Dari only

http://www.bamyantourism.org/index.php?page=en_The+Land [English]
http://www.bamyantourism.org/index.php?page=da_The+Land [Dari]
and other pages (the land, the people, money, when to go, etc.)

http://census.ohio.gov/documents/Questionnaire/2010_Questionnaire_Info.pdf
http://2010.census.gov/2010census/pdf/LAG_Dari.pdf [2 page brochure]

Canada, Settlement.org
six documents in English and Dari
http://www.settlement.org/translatedinfo/
http://www.settlement.org/translatedinfo/resources.asp?t=DARI&l=Dari

www.watchafghanistan.org/newsletter.php
21 issues of a newsletter, English/Dari
note: last section of newsletter contains article summaries, which are not translated

Ministry of Finance
www.mof.gov.af/?lang=en&p=announcements
www.mof.gov.af/?lang=da&p=announcements
some English/Dari parallel articles
note: contents match, but English side uses links to sub-pages

Afghanistan Watch
www.watchafghanistan.org/
mostly English/Dari – scroll through summaries on main page, click on English/Dari links
example:
www.watchafghanistan.org/article025.php [English]
www.watchafghanistan.org/article0025.php [Dari]
example:
www.watchafghanistan.org/article024.php [English]
www.watchafghanistan.org/article0024.php [Dari]
example:
www.watchafghanistan.org/files/The_First_Experience[Voting_Patterns_and_Political_Alignments_in_Wolesi_Jirga(2005-2010)]_English.pdf
www.watchafghanistan.org/files/The_First_Experience[Voting_Patterns_and_Political_Alignments_in_Wolesi_Jirga(2005-2010)]_Dari.pdf

AIHRC – Afghanistan Independent Human Rights Commission
quarterly report
www.aihrc.org.af/2010_eng/Eng_pages/Reports/First%201389%20Quarterly%20Report%20(English).pdf
www.aihrc.org.af/2010_dari/Dri_Pages/Reports/First%201389%20Quarterly%20Report%20(Dari).pdf
press releases
www.aihrc.org.af/2010_eng/Eng_pages/Press_releases/2010/Pre_23_Feb_2010.pdf
www.aihrc.org.af/2010_Dari/Dri_Pages/Press_Releases/2010/pre_dri_23_Feb_2010.pdf

Lesser Resources - English/Pashto only

Department of Health, UK -- translated document list
http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/DH_4123594
Seven brochures in English and Pashto.
Swine flu information sheet for asylum seekers, refugees and other foreign nationals in the UK
A guide to immunisations up to 13 months of age
Tuberculosis - the disease, its treatment and prevention
Go Smokefree campaign website (opens new window)
Error! Hyperlink reference not valid.
Introduction to the National Health Service
Pandemic flu: important information for you and your family (October 2005 edition)
Emergency multilingual phrasebook

Canada, Settlement.org
four documents in English and Pashto
http://www.settlement.org/translatedinfo/
http://www.settlement.org/translatedinfo/resources.asp?t=PASH&l=Pashto

Centre for Addiction and Mental Health
http://www.camh.net/About_Addiction_Mental_Health/Multilingual_Resources/index.html
http://www.problemgambling.ca/EN/AboutGamblingandProblemGambling/Pages/default.aspx
http://www.problemgambling.ca/EN/PGDocuments/Pages/TranslatedResourcesPashto.aspx
problem gambling - many sub-articles in both English and Pashto, not all match

UK Health and Safety Executive
5 brochures in English and Pashto
http://www.hse.gov.uk/languages/pashto/index.htm
for example, "Top tips for ladder and stepladder safety"
http://www.hse.gov.uk/pubns/indg405.pdf [English]

http://www.hse.gov.uk/pubns/pashto/indg405.pdf

Dictionaries and Word Lists

Dictionaries and Word Lists – available online

Wilma Heston, Pashto/English dictionary, 1305 entries.
University of Chicago Digital Dictionaries of South Asian Languages archive
includes example sentences (possibly use as parallel text). includes audio files for words and sentences.
http://dsal.uchicago.edu/dictionaries/heston/index.html

also, older Pashto/English dictionary at same site, Raverty (1867)
http://dsal.uchicago.edu/dictionaries/raverty/

Dictionaries and Word Lists

USAID
Dari/Pashto legal glossary (4,000 terms), mentioned here:
http://afghanistan.usaid.gov/en/Article.644.aspx

Dunwoody Press
PDF version Pashto-English dictionary. 55,000 entries. purchase for $45
http://www.dunwoodypress.com/products/-/300
http://www.dunwoodypress.com/148/PDF/PashtoDictionary_sample.pdf

Dictionary Resources with Known Copyright Problems

Dictionaries
Qamosona; Pashto/English dictionaries, also Pashto/French
www.qamosona.com
"The use of the software for commercial organisations and/or governmental organisations is strictly prohibited. You should contact us for a licence."

glossary of computer terms, English/Pashto
www.afghanai.com/IT/Terms/AfTermsA.htm [blocked on base]
2 lists, one by Microsoft, one by "MOC Afghanistan" -- Microsoft list is "for personal or non-commercial purposes only", other list is Copyright 2007 Afghanai.com, all rights reserved.

Minor Word Lists

DLI Basic Language Survival Guide
http://fieldsupport.dliflc.edu/products/dari/pg_bc_LSK/default.html

http://gloss.dliflc.edu
lists 206 Dari lessons and 154 Pashto lessons, with short glossaries for each lesson.

Named Entity Lists or Documents

IEC – Independent Election Commission
a list of presidential candidates
www.iec.org.af/results_2009/leadingCandidate.html
www.iec.org.af/results_2009/Dari/leadingCandidate.html
www.iec.org.af/results_2009/Pashto/leadingCandidate.html
"Final List of Candidates Wolesi Jirga Election 2010" (Wolesi Jirga = lower house of parliament)
a list of 34 provinces, which is also an index page to lists of candidates
www.iec.org.af/eng/content.php?id=4&cnid=45
www.iec.org.af/dari/content.php?id=4&cnid=45
www.iec.org.af/pashto/content.php?id=4&cnid=45
by clicking on each province, you get a list of each candidate and his/her political party
www.iec.org.af/pdf/final/english/kabul.pdf
www.iec.org.af/pdf/final/kabul.pdf
www.iec.org.af/pdf/final/kabul.pdf
in total, there are 2583 candidates listed

IEC –Media Commission Reports
reports on the amount of coverage given to each candidate; text includes names of candidates, media outlets, locations
www.iec.org.af/eng/content.php?id=6&cnid=56 [list of reports in English]
www.iec.org.af/dari/content.php?id=6&cnid=56 [list of reports in Dari]
www.iec.org.af/pashto/content.php?id=6&cnid=56 [list of reports in Pashto]
note: when using Dari or Pashto index, status bar shows hyperlink in English
for example,
www.iec.org.af/pdf/mediacommission/tenth_report_of_media_monitoring_report.pdf
www.iec.org.af/pdf/mediacommission/tenth_media_monitoring_report_dari.pdf
www.iec.org.af/pdf/mediacommission/tenth_report_%20on_mc_pashto.pdf

IEC – JEMB – Joint Electoral Management Body – Media Commission
reports listing names of candidates, media outlets, locations
www.iec.org.af/jemb.org/media_commission/pdf/media_monitoring_report_24-07_16-08-05%2008_eng.pdf
www.iec.org.af/jemb.org/media_commission/pdf/media_monitoring_report_24-07_16-08-05%2008_dar.pdf
www.iec.org.af/jemb.org/media_commission/pdf/media_monitoring_report_24-07_16-08-05%2008_pas.pdf
another report – no Pashto version
www.iec.org.af/jemb.org/media_commission/pdf/media_monitoring_report_17-30Aug_05-09-05_eng.pdf
www.iec.org.af/jemb.org/media_commission/pdf/media_monitoring_report_17-30Aug_05-09-05_dar.pdf
also, a list of media outlets on pages 3-4 of this document:

www.iec.org.af/jemb.org/media_commission/pdf/media_monitoring_report_17-aug-15-sep_01-10-05_eng.pdf
www.iec.org.af/jemb.org/media_commission/pdf/media_monitoring_report_17-aug-15-sep_01-10-05_dar.pdf
www.iec.org.af/jemb.org/media_commission/pdf/media_monitoring_report_17-aug-15-sep_01-10-05_pas.pdf
list of political parties in 2005 (includes 72 parties and their leaders):
www.iec.org.af/jemb.org/eng/Political/political_update_18june.pdf
www.iec.org.af/jemb.org/dari/Political%20Parties/political_update_18june_dar.pdf
www.iec.org.af/jemb.org/dari/Political%20Parties/political_update_18june_dar.pdf
list of provinces (probably English/Dari)
http://www.iec.org.af/jemb.org/cnlists/final/index.html

www.iec.org.af/pdf/wj_polling_centers/kapisa.pdf
a list of about 5800 polling centers, in English and probably Dari
some quality issues in the English translations, the Kabul centers are badly spelled (masque for mosque, e.g.)

World Bank reports including place names, in English and Dari. (see parallel text entry, above)
for example, this report describes each province:
http://siteresources.worldbank.org/AFGHANISTANEXTN/Resources/305984-1297184305854/ProvBriefsEnglish.pdf
http://siteresources.worldbank.org/AFGHANISTANEXTN/Resources/305984-1297184305854/ProvBriefsDari.pdf

ICTJ - International Center for Transitional Justice
a series of 2-page fact sheets (useful for country names)
(see parallel text entry, above)
www.ictj.org

News Sites with Comparable Text, not Parallel Text

US Embassy Kabul
http://kabul.usembassy.gov/
http://dari.kabul.usembassy.gov/
http://pashto.kabul.usembassy.gov/
has Dari and Pashto sections, but can be hard to match articles
http://kabul.usembassy.gov/silk2.html
http://dari.kabul.usembassy.gov/silk1.html
[no Pashto version]

British Embassy Kabul
http://ukinafghanistan.fco.gov.uk/en
http://ukinafghanistan.fco.gov.uk/dr

http://ukinafghanistan.fco.gov.uk/ps
has Dari and Pashto sections, but can be hard to match articles
http://ukinafghanistan.fco.gov.uk/en/news/?view=News&id=662744782
http://ukinafghanistan.fco.gov.uk/dr/news/?view=News&id=662765882
[no Pashto version]

http://www.afghannews.tv/English.html [blocked]
not named in parallel. archive link seems to be empty.
For example, the following article describes the 9th day of a hunger strike in the Dari and Pashto versions, but the 7th day of the hunger strike in the English version.
http://www.afghannews.tv/Pashto/1390/Mezan/18/La_wolosi_jerga.html
http://www.afghannews.tv/Dari/1390/Mezan/18/Nuhumin_roz.html
http://www.afghannews.tv/English/2011/October/9/Unseated.jpg

Voice of Turkey
English, French, Dari, Pashto, etc.
www.trt-world.com/trtworld/en/news.aspx
www.trtpashto.com/ [blocked]
www.trtdari.com/ [blocked]

Bakhtar News
http://bakhtarnews.com.af

Afghan Islamic Press -- requires subscription --
www.afghanislamicpress.com

Multimedia -English/Dari/Pashto

video from whitehouse.gov
http://www.whitehouse.gov/files/videos/mobile/02-obamaAF-eng.mov
http://www.whitehouse.gov/files/videos/mobile/02-obamaAF-dari.mov
http://www.whitehouse.gov/files/videos/mobile/02-obamaAF-pashtun.mov
"Looking at data on Whitehouse.gov, we don't have a lot of traffic coming from Afghanistan and Pakistan because Internet penetration in the region is relatively low at 2% and 11% respectively. However, mobile penetration is much higher. 52% of the 177 million people in Pakistan have at least 1 mobile device and 30% of the 28.4 million in Afghanistan. Given this trend, we produced short video clips of the President's segment to Afghans and had it dubbed in Arabic, Dari, Pashto, and Urdu in order for them to be distributed locally on mobile devices. Given the small screens on phones, subtitling wasn't an appropriate option. The original version in English is also available."

USAID: one press release with parallel video
http://afghanistan.usaid.gov/en/USAID/Article/2337/USAID_Administrator_Recognizes_Progress_and_Encourages_the_Expansion_of_Mobile_Money_in_Afghanistan
http://afghanistan.usaid.gov/en/USAID/Article/2344 [English]

http://afghanistan.usaid.gov/en/USAID/Article/2342/Mobile_Money_in_Afghanistan_Dari
http://afghanistan.usaid.gov/en/USAID/Article/2343/Mobile_Money_in_Afghanistan_Pashto [videos blocked]

http://www.afghannews.tv/English.html [blocked]
broadcasts alternately in Pashto, Dari, English. probably comparable, not parallel. (see note on text articles, above.)

Multimedia - Language Lessons

http://gloss.dliflc.edu
lists 206 Dari lessons and 154 Pashto lessons.
"Each lesson accompanied by text, glossary, grammar notes and multiple exercises with feedback. Most lessons have an audio component." "Reading and listening lessons are based on authentic materials (articles, TV reports, radio broadcasts, etc.)"

http://www.scola.org/scola/Sample.aspx
SCOLA Insta-Classes, including lessons for Dari and Pashto
Every multimedia Insta-Lesson contains a video clip, a native language transcript, English translation, as well as a quiz and vocabulary list. Can download as a single bilingual pdf.

Joint Language University
http://jlu.wbtrain.com/sumtotal/language/DLI%20basic%20courses/Dari/Books/
Dari-English lessons and vocabulary lists

Indiana U. Center for Languages of the Central Asian Regions (CeLCAR).
Has Pashto textbooks, with audio and video CD. purchase for $95.
http://www.iub.edu/~celcar/language_textbooks/pashto.html

Dunwoody Press, Pashto Newspaper reader (published 1984), 51 newspaper articles, including an English translation, $43. audio also available.
http://www.dunwoodypress.com/products/-/73

Multimedia - English/Dari only

http://langmedia.fivecolleges.edu/culturetalk/afghanistan/index.html
Five College Center for the Study of World Languages
(Amherst, UMass Amherst, Smith, Hampshire, Mt. Holyoke)
International students conduct "Culture Talk" video interviews in their home countries. These videos are stored with transcriptions and English translations.
Afghanistan has 45 topic areas, some with multiple videos per topic (e.g., music, politics, education, cooking). For example, here is a video about traditional Afghan dress, and a document containing the transcript in Dari and an English translation.
http://langmedia.fivecolleges.edu/culturetalk/afghanistan/content/af_afghandress_s1e.rm
http://langmedia.fivecolleges.edu/culturetalk/afghanistan/content/af_afghandress_s1e.mp4

http://langmedia.fivecolleges.edu/culturetalk/afghanistan/content/af_afghandress_s1e.doc

Settlement.org, one video in English and Dari
http://www.settlement.org/translatedinfo/
http://www.settlement.org/translatedinfo/resources.asp?t=DARI&l=Dari

Multimedia - English/Pashto only

HRIS – Health Rights Information Scotland
Health care for asylum seekers and refugees in Scotland
http://www.hris.org.uk/media/audio/recordsV4/HowToSeeEntire.mp3 [English]
http://www.hris.org.uk/media/audio/pashtoV3/Pashto%20Leaflet%20Whole.mp3
[Pashto]
http://www.hris.org.uk/resources/asylum-pashto/ [text: English/Pashto]

Wilma Heston, Pashto/English dictionary, 1305 entries.
University of Chicago Digital Dictionaries of South Asian Languages archive
includes example sentences in English and Pashto, with audio files for Pashto words and sentences.
http://dsal.uchicago.edu/dictionaries/heston/index.html

Multimedia – Dari/Pashto only

IEC – Independent Election Commission
radio public service announcements in Dari and Pashto
www.iec.org.af/jemb.org/eng/psa.html [index]
for example,
http://www.iec.org.af/jemb.org/eng/Public%20Information/PSA's/01Track.wma
http://www.iec.org.af/jemb.org/eng/Public%20Information/PSA's/02Track.wma
there are also tv announcements, not clear which language is used
for example,
www.iec.org.af/jemb.org/eng/TV%20Spots/TV%20Spot%201.wmv

Multimedia - Dari only

Five College Center for the Study of World Languages
Dari Audio Samples (not parallel)
34 topics, for example:
http://langmedia.fivecolleges.edu/persian/dari/topics/da_famouscities.mp3

audio recordings of mental health information
http://www.mind.org.uk/help/diagnoses_and_conditions/audio_translations/afghan_dari

Sites to Watch: organizations that may post more resources in the future

Indiana U. Center for Languages of the Central Asian Regions (CeLCAR). "the critical languages of Central Asia (and Afghanistan and Pakistan)"... "meeting strategic national needs"
http://www.indiana.edu/~celcar/

CAL Pashto-English Glossary, 5,000 entries. --scanned document only--
http://www.eric.ed.gov/PDFS/ED364083.pdf

U. Minnesota, Digital South-Asia Language Archive ("the languages of Afghanistan, India and Pakistan") currently has scanned documents, transliterated.
http://lrc.lib.umn.edu/dsala.htm

PAN Localization project www.panl10n.net "promoting language technology across developing Asia" in collaboration with IDRC, Canada www.idrc.ca, National U. of Computer and Emerging Sciences, Pakistan www.nu.edu.pk

Wilma Heston dictionary mentioned in dictionary section, plans for expansion
A digital Pashto-English dictionary with audio : 1,000 words from a core vocabulary
http://dsal.uchicago.edu/dictionaries/heston/index.html

World Health Organization, mentions Dari and Pashto translations of some documents, not available online
http://www.who.int/surgery/activities/en/
http://www.who.int/patientsafety/research/methods_measures/human_factors/individual_tools/en/

Afghanistan Supreme Court: lists online magazines, with links set up for English, Dari, Pashto, but so far these appear to only be in Dari. check to see if they post translations in the future.
http://supremecourt.gov.af/en/documents?DID=127

APPENDIX B - Farsi and English Parallel Text Resources

Farsi-English Parallel Text Resources

We have identified these main resources for Farsi/English parallel text:

1. IIP Digital http://iipdigital.usembassy.gov
2. Virtual US Embassy to Iran http://iran.usembassy.gov/
3. US Centcom http://www.centcom.mil/ur
4. Health, Immigration, and Various Minor Resources

1. IIP Digital
http://iipdigital.usembassy.gov

News articles and longer publications translated into several languages. Also available are English-language multimedia, with transcripts in English and in Farsi.

The parallel pages are not named in parallel, but the site provides links that can be used to collect the corresponding articles. Only some of the material is translated into Farsi. In order to collect just those articles that have Farsi translations, go to the Farsi index:
http://iipdigital.usembassy.gov/iipdigital-fa/index.html#axzz1xbbifdmf

2. Virtual US Embassy to Iran

http://iran.usembassy.gov/
This site contains articles named in parallel (with the word "persian" before the rest of the address). Of particular interest are the sections titled News & US Policy, and Open Societies

News & US Policy (News Domain)
http://iran.usembassy.gov/news-policy.html
For example,
http://iran.usembassy.gov/sweden.html
http://persian.iran.usembassy.gov/sweden.html

Open Societies (General Interest Domain)

http://iran.usembassy.gov/open-societies.html

Note: Many of the articles in this section are links to the IIP site, which we will collect separately.

3. US Centcom http://www.centcom.mil/fa
News articles translated into English, Urdu, Farsi, Russian, and Arabic; named in parallel
18 pages of listings of news articles with translations (about 11 per page, = about 200 articles)
oldest translated article: Jan. 2010
example of naming: [generally, can replace ur (Urdu) with fa (Farsi) in these examples]
http://www.centcom.mil/news/mine-dog-teams-get-ready-for-action
http://www.centcom.mil/ur/news/mine-dog-teams-get-ready-for-action

note: a few, most recent, articles not translated; an English-text placeholder is used until they are translated.

There are also press releases: (7+ pages of listings, approx 80 articles)
http://www.centcom.mil/press-releases/cmf-responds-to-medical-emergency-aboard-iranian-fishing-vessel
http://www.centcom.mil/ur/press-releases/cmf-responds-to-medical-emergency-aboard-iranian-fishing-vessel

There is also a list of 20 countries:
http://www.centcom.mil/fa/area-of-responsibility-countries
with a page for each country:
http://www.centcom.mil/afghanistan/
http://www.centcom.mil/fa/afghanistan/

4. Health, Immigration, and Various Minor Resources
Ontario immigration information http://www.ontarioimmigration.ca
Improving Your Language Skills
Opportunities Ontario
Opportunities Ontario Students
Services to Help You Settle
Welcome Letter from the Premier
Work in Your Profession

settlement.org, 53 documents, may overlap with Ontario information, above
http://www.settlement.org/translatedinfo/resources.asp?t=FARSI&l=Farsi

http://healthtranslations.vic.gov.au/bhcv2/bhcht.nsf/PresentMultilingualResource?Open&x=&s=Farsi_(Persian)   271 articles, listed by English titles; click on title to go to page listing all available translations.

Victoria, health articles, 271 listed for Farsi http://www.healthtranslations.vic.gov.au
http://www.healthtranslations.vic.gov.au/bhcv2/bhcht.nsf/PresentMultilingualResource?Open&x=&s=Farsi_(Persian)

World Bank, 12 documents
http://www-wds.worldbank.org/external/default/main?menuPK=64258548&pagePK=64187838&piPK=64187928&theSitePK=523679&function=BrowseFR&menuPK=64258548&siteName=WDS&conceptattcode=Farsi~434541&pathtreeid=LANG&sortattcode=DOCDT+Desc

APPENDIX C - Online Resources for African Languages

Online Resources for African Languages:
Electronic Resources for Machine Translation of Yoruba, Hausa, Igbo, Somali, and Swahili

--draft-- October 2010

This document lists electronic resources that are available for general use.

Resources are classified as:

Parallel Text
Dictionaries
Monolingual Text
Other

Dictionaries are further classified as:

Text: can download entire text of dictionary
ABC: one page per letter, can download one page at a time
Online: a search interface, difficult to access word list

"Other" resources include multimedia, software such as spell-checkers, projects under development, and proprietary resources.

Yoruba

Parallel Text

Parallel Text; 7 documents Yoruba/English
Refugee Health Information Network
http://rhin.org/search/search_results.asp?quick_search=&language=79&Image1.x=37&Image1.y=13

Dictionaries

Text Dictionary; 368,000 entries; Yoruba-English and English-Yoruba; includes POS
LDC: Global Yoruba Lexical Database v.1.0 (LDC2008L03)

ABC Dictionary; __ entries; English-Yoruba only
http://www.yorubadictionary.com/index.html

Online Dictionary; __ entries; English-Yoruba, Yoruba-English; includes POS
http://words.fienipa.com (drop-down language menu)

Monolingual Text

Monolingual Text; An Crubadan web crawl; 871,445 words
corpus statistics available
http://borel.slu.edu/crubadan/index.html

Other

list of common Yoruba names; __ entries
http://www.yorubadictionary.com/yorubanames.htm

African Languages Technology Initiative (Alt-i)
plans for Yoruba ASR using tones; Yoruba TTS; Yoruba-English MT; Yoruba spell checker;
www.alt-i.org
www.aclweb.org/anthology/W/W09/W09-0708.pdf

A computational approach to Yoruba morphology; Finkel and Ajadi, 2009
http://www.aclweb.org/anthology/W/W09/W09-0704.pdf

Hausa

Parallel Text

Parallel Text; approx. 1500 sentences; Hausa/English
Hausar Baka Corpus [video transcripts]
(50 videos with Hausa transcripts, of which 19 have English translations; primarily simple text translated so far)
http://www.humnet.ucla.edu/humnet/aflang/hausarbaka/

Parallel Text; 767 sentences; Hausa/English
Novel, Ruwan Bagaja, by Abubakar Imam
www.abubakarimam.com (Hausa)
http://african.lss.wisc.edu/hunter/306/ruwan.htm (English)

BBC
http://www.bbc.co.uk/hausa/

Dictionaries

Text Dictionary; __ entries; Hausa-English; includes POS
(all the words from the Hausar Baka corpus)
http://www.humnet.ucla.edu/humnet/aflang/hausarbaka/Download_vocabulary.html

Text Dictionary; 16,915 entries; Hausa – German,English; includes POS
K'ofar Hausa wordlist
http://www.univie.ac.at/Hausa/KamusTDC/CD-ROMHausa/KamusTDC/ARBEIT2.txt

Online Dictionary; __ entries; Hausa-English; includes POS
http://words.fienipa.com (drop-down language menu)

Online Dictionary;39,000 words; Hausa-English and English-Hausa (4600 words); includes POS
On-Line Bargery Dictionary
copyright issues? Nakamura Hriokazu Sule, Japan
http://maguzawa.dyndns.ws

Monolingual Text

Monolingual Text; An Crubadan web crawl; 6,164,518 words
corpus statistics available
http://borel.slu.edu/crubadan/index.html

http://www.voanews.com/hausa/news/ (Voice of America)
http://www.dw-world.de/dw/0,,627,00.html (Deutche Welle)
http://hausa.cri.cn/ (China Radio International)

Corpus; 34,320 sentences from 2321 documents (not available yet?)
Hausa Internet Corpus, VOA Corpus (VOA news 2001-2009)
http://www.sfb632.uni-potsdam.de/hausa/resources/news-corpora/voa-corpus

Other
video: Hausar Baka corpus, 5 hours
http://www.humnet.ucla.edu/humnet/aflang/hausarbaka/

annotation: Hausar Baka corpus annotated for POS, morphology, syntax, etc.
(not yet available?)
Hausa Internet Corpus, Annotations of the Hausar Baka Corpus
http://www.sfb632.uni-potsdam.de/hausa/resources/annotation/annotations-hausar-baka-corpus

Word List; 15,000 entries; Hausa only (no glosses); gives POS and morphological info
(includes info from Bargery and K'ofar Hausa dictionaries)
Hausa Internet Corpus, Morphosyntactic Wordlist
http://www.sfb632.uni-potsdam.de/hausa/resources/tools/morphosyntactic-wordlist

Comments
Hausa is written in both a Roman-based alphabet and an Arabic script, called Ajami. Resources listed here are in the Roman alphabet.

Igbo
Parallel Text
Parallel Text; over 1000 sentences; Igbo and English
includes greetings, questions and answers, proverbs; includes audio
Igbo911
www.igbo911.com/index.htm

Parallel Text; 7 documents; Igbo and English
Refugee Health Information Network
http://rhin.org/search/search_results.asp?quick_search=&language=32&Image1.x=33&Image1.y=14

Parallel Text; __ sentences; Igbo and English
includes dialogs and vocabulary lists
Igbo Insight Guide
www.igboguide.org

Dictionaries

Text Dictionary; approx 13,000 entries? Igbo-English; includes POS
Williamson/Blench, Dictionary of Onicha Igbo
http://www.rogerblench.info/Language%20data/Niger-Congo/Benue%20Congo%20West/Igboid/IGBO%20Dictionary.pdf

Text Dictionary; 6602 entries; Igbo-English
African Language Research Project, U. Maryland Eastern Shore
www.umes.edu/alp/lexicon1/public/resultsb.asp

Text Dictionary; about 400 entries; Igbo-English and English-Igbo
Igbo Insight Guide
http://www.igboguide.org/index.php?l=vocabulary

Monolingual Text

Monolingual Text; An Crubadan web crawl; 397,661 words
corpus statistics available
http://borel.slu.edu/crubadan/index.html

Other

Igbo OCR system
Okan Kolak & Philip Resnik, OCR Post-Processing for Low Density Languages, ACL-HLT (2005).
http://www.aclweb.org/anthology/H/H05/H05-1109.pdf

Igbo Archival Dictionary Project, Igbo Online Resources Project
plans to include spell checkers, speech synthesis, tone-marking, syllabification, machine translation, dialect maps
Wanjiku Ng'ang'a, Towards a Comprehensive, Machine-readable Dialectal Dictionary of Igbo, AFLAT 2010
www.aflat.org/files/nganga.pdf

African Language Research Project, U. Maryland Eastern Shore
plans for creation of electronic databases and machine translation
www.umes.edu/alp

Possible parallel audio from Voice of Nigeria radio broadcasts www.voiceofnigeria.org
some parallel VON English/French/Igbo broadcasts listed on a school website:
www.igboschool.com/Igbo_Radio.php

African Languages Technologi Initiative (Alt-i)
plans for Igbo-English MT
www.alt-i.org

Igbo911 emphasis on dialects, plans to create dialectical lexicon, text-to-speech projects
Dr. Izu Asiegbunam, Toronto, Canada

www.igbo911.com/index.htm

Comments
Some Igbo people work to preserve the language, but others prefer to teach their children only English. This attitude may limit the amount of material posted electronically in Igbo.

Somali
Parallel Text
Parallel Text; approx. 17,500 sentences in 285 documents; Somali- English
Refugee Health Information Network
http://rhin.org/search/search_results.asp?quick_search=&language=62&x=40&y=8

Parallel Text; 89 documents; Somali-English; includes audio, video
US National Library of Medicine
http://www.healthyroadsmedia.org/somali/index.htm

Parallel Text; 24 documents; Somali-English
Unicef newsletter, press releases
http://www.unicef.org/somalia/media_187.html (English)
http://www.unicef.org/somalia/media_4738.html (Somali)
http://www.unicef.org/somalia/media_65.html (English)
http://www.unicef.org/somalia/media_4879.html (Somali)

Parallel Text; 15 documents; Somali-English
Ethnomed, University of Washington Health Sciences Libraries
http://ethnomed.org/patient-education/somali

Parallel Text; 8 documents; Somali-English
Virginia Department of Health
http://www.vdh.virginia.gov/CLAS_Act/languageresources/translatedvdh/lang_somali.htm

BBC
http://www.bbc.co.uk/somali/

Dictionaries
ABC Dictionary; __ entries; Somali-English and English-Somali
http://markacadey.banadir24.com/go/dictionary/

Online Dictionary; claims 12,331 entries; Somali-English-Italian
some entries have no glosses, not clear how many are translated
http://www.redsea-online.com/modules.php?name=dictionary

Monolingual Text
Monolingual Text; An Crubadan web crawl; 4,256,336 words
corpus statistics available

http://borel.slu.edu/crubadan/index.html

http://www.voanews.com/somali/news/ (Voice of America)

Other
video; 80 recordings; Somali and English
Refugee Health Information Network
http://rhin.org/search/search_results.asp?quick_search=&language=62&x=40&y=8

Online Spell Checker; 108,371 words; Somali
http://www.redsea-online.com/modules.php?name=dictionary&func=spellcheck

MT Tools: automatic speech to text transcription for Somali; morphological analyzer; monolingual corpus/language model
Automatic transcription of Somali language, Abdillahi Nimaan, et al., Interspeech 2006.

Swahili

Parallel Text
Annotated Text; 12.5 million words; annotated with English gloss, POS, morphology
licensed for academic research only
Helsinki Corpus of Swahili
www.aakkl.helsinki.fi/cameel/corpus/intro.htm

Parallel Text; 34 documents; Swahili/English
Refugee Health Information Network
http://rhin.org/search/search_results.asp?quick_search=&language=65&Image1.x=28&Image1.y=14

BBC
http://www.bbc.co.uk/swahili/

Channel Africa (not sure if the articles are parallel)
broadcasts in Chichewa, Silozi, Swahili, English, French and Portuguese
www.channelafrica.co.za

Dictionaries
Text Dictionary; 61,087 entries; Swahili-English and English-Swahili; includes POS
Kamusi Project
http://www.kamusi.org/?q=en/dictionaries

Text Dictionary; 58,000 entries; English-Swahili
available for purchase for 45 EUR
European Language Resources Association, English => Swahili Bilingual Lexicon
http://catalog.elra.info/product_info.php?products_id=1060

Online Dictionary; 16,000 entries; Swahili-English and English-Swahili; includes POS
Download available for $11
TshwanDJe Swahili-English Dictionary
http://africanlanguages.com/swahili/index.php?|=en

Text Dictionary; 1400 entries; English-Swahili
Comparative Bantu OnLine Dictionary (CBOLD)
(full database has 200 languages, 445,00 entries)
http://www.cbold.ish-lyon.cnrs.fr/

Online Dictionary; 2600 entries; Swahili-English
FreeDict Swahili-English Dictionary
www.freedict.org

Online Dictionary; __ entries; Swahili-English; includes POS
http://words.fienipa.com (drop-down language menu)

Monolingual Text
Monolingual Text; An Crubadan web crawl; 5,165,051words
corpus statistics available
http://borel.slu.edu/crubadan/index.html

http://www.voanews.com/swahili/news/ (Voice of America)
http://www.dw-world.de/dw/0,,633,00.html?id=633 (Deutche Welle)
http://swahili.cri.cn/ (China Radio International)

Other
Parallel Text; approx half a million words; Swahili/English
SAWA Corpus; includes Bible, Quran; online Kamusi dictionary examples; movie subtitles
not available ?
www.aclweb.org/anthology/W09-0702

Swahili Language Manager (SALAMA); morphological analysis, syntactic analysis, semantic disambiguation, translation from Swahili to English
http://www.njas.helsinki.fi/salama/

Swahili spellchecker; 70,000 Swahili words; for use with Jambo OpenOffice
http://www.kamusiproject.org/sw/software

Swahili POS tagger; Guy De Pauw, et al., Data-Driven Part-of-SpeechTagging of Kiswahili, Text, Speech, and Dialog 2006; POS tagger based on the Helsinki Corpus
http://www.cnts.ua.ac.be/Publications/2006/De 06c/

Rule-based Swahili to English Machine Translation; Academy of Finland
http://www.njas.helsinki.fi/salama/translat.html

APPENDIX D - Problems with the LDC Urdu Language Pack

Problems and Approaches for the LCTL_Urdu_v1.0 Language Pack

DRAFT: 14 May 2010

Introduction

This document outlines the problems researchers in the AFRL SCREAM laboratory discovered while working with the LDC's LCTL Urdu language pack, LCTL_Urdu_v1.0. The document is organized into three main parts: Part A describes errors we found in the LCTL materials; Part B notes general characteristics of the Urdu language that cause difficulties; and Part C outlines some of the approaches that we found useful to address these problems in the SCREAM lab.

Note: MS Word tends to reverse the Arabic script examples when the entire line is in Arabic script. For example, the order of typing of the following words is "staff reporter", but MS Word displays them as "reporter staff" (reading from right to left):
رپورٹر اسٹاف
The "staff reporter" order can be verified by looking at the order of the Unicode codepoints:

0627	0633	0679	0627	0641	0020	0631	067E	0648	0631	0679	0631
a	s	t	a	f	[space]	r	p	w	r	t	r

For this reason, it may be necessary to view this document in a different program, such as Notepad.

A. Problems found in the LDC Urdu language pack

A.I. lexicon

A.I.a. reversals

A.I.a.i. left-to-right ordering error in acronyms.

for example, here we see "com dot Urdu C B B" -- each word is run right-to-left, but the phrase is reversed.

```
<ENTRY id="LEX-URD-00852000">
      <WORD>کام ڈاٹ اردو سی بی بی</WORD>
      <STEM>کام ڈاٹ اردو سی بی بی</STEM>
      <MORPH>noun</MORPH>
      <POS>NOUN</POS><GLOSS source="dictionary.xml">BBC Urdu dot Com</GLOSS>
   </ENTRY>
```

compare above with a correct entry,

```
<ENTRY id="LEX-URD-00947500">
      <WORD>بی بی سی</WORD>
      <STEM>بی بی سی</STEM>
      <MORPH>noun</MORPH>
      <POS>NOUN</POS><GLOSS source="dictionary.xml">British Broadcasting
Corporation</GLOSS>
   </ENTRY>
```

(note also inconsistency, acronyms are sometimes translated by letters, sometimes by words)

A.I.a.ii. left-to-right ordering error in phrases.

There are dozens of phrases in which the order of the words has been reversed.
for example, we see here "(LRK) lybartry rysrch khuta" for "Kahuta Research Laboratory (KRL)"

```
   <ENTRY id="LEX-URD-00918900">
      <WORD>ایل آر کے) لیبارٹری ریسرچ کہوٹہ(</WORD>
      <STEM>ایل آر کے) لیبارٹری ریسرچ کہوٹہ(</STEM>
      <MORPH>noun</MORPH>
      <POS>NOUN</POS><GLOSS source="dictionary.xml">Kahuta Research Laboratory
(KRL)</GLOSS>
   </ENTRY>
```

And here we find "paol koln farxh ozir amriki", meaning "Powell Colin foreign-affairs minister American", a reversal of "American Foreign Secretary Colin Powell"

```
   <ENTRY id="LEX-URD-02454000">
      <WORD>پاول کولن خارجہ وزیرِ امریکی</WORD>
      <STEM>پاول کولن خارجہ وزیرِ امریکی</STEM>
      <MORPH>noun</MORPH>
      <POS>NOUN</POS><GLOSS source="dictionary.xml">American Foreign Secretary
Colin Paul</GLOSS>
   </ENTRY>
```

In some cases, the lexicon already contains the entry in the correct order, as in these two entries for Isalm-abad:

abad - islam

```
   <ENTRY id="LEX-URD-00235300">
      <WORD>آباد اسلام</WORD>
      <STEM>آباد اسلام</STEM>
      <MORPH>noun</MORPH>
      <POS>NOUN</POS><GLOSS source="dictionary.xml">Islamabad</GLOSS>
   </ENTRY>
```

islam - abad

```
   <ENTRY id="LEX-URD-00729200">
      <WORD>اسلام آباد</WORD>
      <STEM>اسلام آباد</STEM>
      <MORPH>noun</MORPH>
      <POS>NOUN</POS><GLOSS source="dictionary.xml">Islamabad</GLOSS>
   </ENTRY>
```

But in many cases, we only have the reversed entry.

A.I.b. comma-separated glosses

Many entries with multiple glosses do not list the glosses with separate <GLOSS> tags, but rather as a comma-separated list. We have had to edit these into separate GLOSS entries in order to be able to look up POS tags for factored language processing.

Separate <GLOSS> listings:

```
<ENTRY id="LEX-URD-00003500">
      <WORD>کلیان</WORD>
      <POS>NONE</POS><GLOSS source="urdu.lex">benediction</GLOSS>
      <POS>NONE</POS><GLOSS source="urdu.lex">blessedness</GLOSS>
      <POS>NONE</POS><GLOSS source="urdu.lex">felicity</GLOSS>
   </ENTRY>
```

Combined <GLOSS> list:

```
<ENTRY id="LEX-URD-01531800">
      <WORD>کوچ</WORD>
      <STEM>کوچ</STEM>
      <MORPH>Noun</MORPH>
      <POS>NOUN</POS><GLOSS
source="dictionary.xml">death,decease,deprature,gait,march,remove,route,sofa</GLOSS>
   </ENTRY>
```

A.I.c. errors in part of speech tags
A.I.c.i. erroneous pos tags
For example, “attain” as a noun, “thirty” as a verb.

A.I.c.ii. lack of POS information
More than 18,000 entries have POS = “NONE”

A.I.c.iii. inconsistent POS tags
15 entries with POS listed as "OTHER"
1 entry with POS listed as “ADJ or ADV” (entry 00623700)

A.I.d. problems with morphological information
A.I.d.i. MORPH elements not connected to GLOSS elements
Inflectional information is represented with separate <MORPH> tags; when there are multiple glosses, there is no principled way to connect the <MORPH> information with the appropriate <POS>, <GLOSS> listings. For example,

```
   <ENTRY id="LEX-URD-00004100">
      <WORD>تبدیلی</WORD>
      <STEM>تبدیل</STEM>
      <MORPH>Noun+Fem+Sg</MORPH>
      <MORPH>Noun+NameofAct</MORPH>
      <MORPH>Adjective+Fem</MORPH>
      <MORPH>noun</MORPH>
      <MORPH>Noun+Dim</MORPH>
```

```
        <POS>NOUN</POS><GLOSS source="dictionary.xml">Alteration,Change,Revision,Variance,removal,remove,shift,substitution,turn</GLOSS>
        <POS>ADJ</POS><GLOSS source="dictionary.xml">Alteration,Change,Revision,Variance,removal,remove,shift,substitution,turn</GLOSS>
        <POS>NOUN</POS><GLOSS source="dictionary.xml">alteration,removal</GLOSS>
        <POS>NOUN</POS><GLOSS source="dictionary.xml">Alteration,Change,Revision,Variance,removal,remove,shift,substitution,turn</GLOSS>
    </ENTRY>
```

A.I.d.ii. Inconsistencies in the MORPH tags

The <MORPH> tags are inconsistent. We see:

Masc	masc		
Noun	noun	nounnoun	nounverb
Dim	dim	dimnoun	
NameofAct		nameofact	
Sg	sg	sgnoun	
Pl	Plu	pl	
Verb	verb		
Hon	honverb		
Adjective		Adjectiveective	

There are apparent redundancies in the lists of inflectional features:
Adjective+masc
Adjective+mascadj+masc
verb+directcaus+nonpast+p+sgverb+directcaus+nonpast+p+sgverb+directcaus+past+masc+pl

We also see <MORPH>NONE</MOPRH>.

A.I.e. lack of entries for inflected forms

When looking up words from the parallel text in the lexicon, words that show up frequently as unknowns include inflected forms of common verbs like "go", "do", "have", as well as plurals and oblique forms of common nouns.

Consider the paradigm for masculine nouns:

nominative singular	–a	nominative plural	–e
oblique singular	–e	oblique plural	–õ

For some nouns, the lexicon only has an entry for the nominative singular. For example, the lexicon has one entry for fisherman in the nominative singular, /mchhyra/ but the parallel text also includes forms for fishermen, oblique singular, /mchhyre/ and fishermen, oblique plural, /mchhroN/.

nominative singular in –a: مچھیرا

```
<ENTRY id="LEX-URD-00855500">
   <WORD>مچھیرا</WORD>
   <STEM>مچھیرا</STEM>
   <MORPH>Noun</MORPH>
   <POS>NOUN</POS><GLOSS source="dictionary.xml">fisher,fisherman</GLOSS>
</ENTRY>
```

oblique singular in –e مچھیرے

VOA_URD_20060410.5.ltf.xml
احتجاج باقاعدہ سے بھارت کا پاکستان پر ہلاکت کی مچھیرے پاکستانی ایک ہاتھوں کے گارڈ کوسٹل بھارتی
Pakistan Protests to India over Indian Coast Guard Killing Pakistani Fisherman

oblique plural in –õ مچھیروں

VOA_URD_20060127.4
ہے لیا کر گرفتار پر بناء کی ورزی خلاف سرحدی پر طور مبینہ کو مچھیروں بھارتی 37نے پاکستان
Pakistan Detains 37 Indian Fishermen for Alleged Border Violation
For other nouns, the lexicon has the nominative forms for both singular and plural, but lacks the oblique forms. So, for the noun ‘department, institution, organization, agency’ we find /adarh/ nominative singular, and /adare/ nominative plural, but not /adaroN/ oblique plural.

nominative singular in –h ادارہ

```
<ENTRY id="LEX-URD-01813300">
   <WORD>ادارہ</WORD>
   <STEM>ادارہ</STEM>
   <MORPH>Noun</MORPH>
   <POS>NOUN</POS><GLOSS
source="dictionary.xml">Organization,institute,institution</GLOSS>
</ENTRY>
```

nominative plural in –e ادارے

```
<ENTRY id="LEX-URD-02210500">
   <WORD>ادارے</WORD>
   <STEM>ادارے</STEM>
   <MORPH>noun</MORPH>
   <POS>NOUN</POS><GLOSS source="dictionary.xml">Departments</GLOSS>
</ENTRY>
```

oblique plural in –õ اداروں

JANG_URD_20060331.76850

برطانوی ماہرین نے اتوارکو اسلام آباد میں وزارت داخلہ کے حکام کے علاوہ پولیس اور دیگر سیکورٹی اداروں کے اہم
افراد کے ساتھ ملاقات کی، جس میں سیکورٹی پلان کا ازسرنو جائزہ لیا گیا۔
British security experts hold talks with the police and other important people of other security agencies apart from the officials of the Interior ministry in Islamabad in which the security plans were discussed afresh.

A.I.f. spelling errors in the Urdu word entries
A.I.f.i. misspelled Urdu words

For example:
New Delhi spelled with a w instead of a d:

```
  <ENTRY id="LEX-URD-00555200">
    <WORD>نئ وہلی</WORD>
    <STEM>نئ وہلی</STEM>
    <MORPH>noun</MORPH>
    <POS>NOUN</POS><GLOSS source="dictionary.xml">New Delhi</GLOSS>
  </ENTRY>
```

New Delhi spelled as one word (also reversed):

```
  <ENTRY id="LEX-URD-01031000">
    <WORD>دہلینئی</WORD>
    <STEM>دہلینئی</STEM>
    <MORPH>noun</MORPH>
    <POS>NOUN</POS><GLOSS source="dictionary.xml">New Delhi</GLOSS>
  </ENTRY>
```

A.I.f.ii. spelling errors involving duplicate diacritics

For example, entry 02350900 has 6 instances of the diacritic, 064F, Arabic Damma.

```
  <ENTRY id="LEX-URD-02350900">
    <WORD>اُُُُُُترنا</WORD>
    <POS>LOC</POS><GLOSS source="urdu.lex">down</GLOSS>
  </ENTRY>
```

If the word processor displays diacritics as an overstrike, the extra characters may not be visually distinguished, but they will still cause the word to be considered as a distinct form for computation.

We found duplicate diacritics in the following lexicon entries:

00226900 00695700 00855600 01147300 01531700
01767100 01837900 02175000 02350900 02406200

A.I.g. problems with the glosses
A.I.g.i. spelling errors

There are many misspelled glosses: wickwets, Valter, e4ngross, Saints Peter Berg, gneius, theif, strenght, reprtition, scrap-bookm, explaination, etc.

A.I.g.ii. rare meanings

The glosses include rare meanings like “manubrium” for “handle” (LEX-URD-01454900) or “fistula” for “reed” (LEX-URD-0198100). Also, the rare meanings are sometimes listed first, which is not helpful if we try to use the dictionary for practical purposes such as the translation of persistent unknowns. For example, the word نے /ne/ is most commonly used in this data as the ergative postposition, but can also refer to a kind of flute, a ney. The first meaning listed is for this word is “fistula”; the possible ergative meaning, “by” is listed last. (We also see here an example where the POS and MORPH tags seem to be inconsistent with the meanings.)

```
- <ENTRY id="LEX-URD-00856200">
  <WORD>نے</WORD>
  <STEM>دے</STEM>
  <MORPH>Verb+Inf+Pl</MORPH>
  <POS>VERB</POS>
  <GLOSS
source="dictionary.xml">fistula,flageolet,flute,hauntboy,pipe,reed,syphon,tube,by</GLOSS>
  </ENTRY>
```

A.I.g.iii. additional information within the glosses

There are also some parenthetical notes within translations. e.g.: "Went (pl.)" and "than (comparative)" These can be a problem if using the dictionary to translate certain words directly.

A.I.h. omissions

The lexicon has uneven coverage of proper nouns. For example, Karachi is an unknown word (although there is an entry for Karachi Press). Also, the dictionary lacks certain domain-specific vocabulary: In particular, the parallel text contains several Urdu words for cricket terminology that are not found in the lexicon.

A.I.i. interaction with spelling normalization

Some words in the parallel text cannot be found in the dictionary because of the use of alternate characters, such as Arabic kaf instead of the Urdu kaf used in the lexicon. (This is not always obvious when looking at the words in a word processor, since the medial forms may look the same.) Spelling normalization of the text generally helps, but also creates problems for other words, such as some Arabic names, which are entered in the lexicon in Arabic spelling. Thus, any spelling normalization effort needs to consider both the data and the lexicon.

Some examples where the parallel text differs from the lexicon:

The lexicon entry for the word میں “in” or “I” is spelled with the Urdu vowel 06CC, but the data contain some sentences in which this word is spelled with the Arabic vowel, 064A. (the medial forms appear the same)

ں ی م = میں

```
  <ENTRY id="LEX-URD-01624300">
```

```
      <WORD>میں</WORD>
      <POS>NONE</POS><GLOSS source="urdu.lex">I</GLOSS>
      <POS>NONE</POS><GLOSS source="urdu.lex">Into</GLOSS>
      <POS>NONE</POS><GLOSS source="urdu.lex">about</GLOSS>
      <POS>NONE</POS><GLOSS source="urdu.lex">amid</GLOSS>
....
   </ENTRY>
```

ں ي م = میں
MEHR_URD_20060331.99818
:میں موصل شہر کے عراق
In the city of Mosul, Iraq

We also see some data in which the Arabic kaf ك is used instead of the Urdu kaf ک , as in the word American ي ك ي ر م ا = امريكي below:

MEHR_ENG_20060331.97859
The Mehr News Agency has reported, with information from Al-Arabia TV, that it appears American soldiers want to arrest Muqtada al-Sadr.
خبر رساں ايجنسي مہر نے العربيہ ٹي وي کے حوالے سے خبر دي ہے کہ ايسا محسوس ہوتا ہے کہ امريكي فوجي مقتدي
صدر کو گرفتار کرنا چاہتے ہیں

On the other hand, some Arabic names like "al Hasani" are found in the lexicon with the Arabic spelling, ي Unicode 064A; this matches the spelling found in the text. In this case, normalization of 064A to 06CC in the text would prevent us from accessing the lexical entry.

```
   <ENTRY id="LEX-URD-00864200">
      <WORD>الحسني</WORD>
      <STEM>الحسني</STEM>
      <MORPH>noun</MORPH>
      <POS>NOUN</POS><GLOSS source="dictionary.xml">alHasani</GLOSS>
   </ENTRY>
```

CRL_ENG_20060119.9
Mr. Hajim aL-Hassani has strongly opposed the attack on Fallujah.
تھے۔ مخالف شدید بھی کے حملے پر فالوجہ کے فوج امریکی الحسني حاجم

A. II. spelling normalization

A.II.a. errors in script, Encoding_Conversion/Software/UrduPresentation2Regular.tcl

remove:	add:
ff91 > 06a9	fb91 > 06a9
fed0 > 6044	fedd > 0644
fea6 > 06c1	
fea7 > 06c1	

fea8 > 06c1
fea9 > 06c1

explanations:
typo: ff91 to 06a9, should be fb91 to 06a9 (ff91 is a katakana character)
double typo: fed0 > 6044 (a Chinese character) should probably be fedd > 0644 (L to L)
(note that fed0 is already correctly mapped to 063a)
variant forms: fea6, fea7, fea8 were all mapped twice, to 062e (correct h) and to 06c1 (incorrect h, per Unicode charts)
fea9 got swept up in this, perhaps, and maps to 062f (correct d), and to 06C1 (incorrect and implausible)

A.II.b. omissions from UrduPresentation2Regular.tcl
The parallel text also includes Arabic presentation characters (FB50-FDFF, FE70-FEFF) which are not included in the mapping in UrduPresentation2Regular.tcl. For example, we find Arabic presentation forms here:

UDI_URD_20060222.visa
اق ےنپراذ یبیروجینافسر ناختہ ےس رابطہ کر ےتکس آپ وت ںیہ ےتہاچ انرک لصاح تامولعم دیزم قلعتم ےک ازیو آپ رگا
یہں-
A visa is a document showing that you have leave to enter Norway and other Schengen countries for a limited period of up to 90 days.

There are also some variant Urdu characters which are not included in the mapping.

Current		Needed	
ك	0643	ک	06A9
ي	064A	ی	06CC
ٓا	0653-0627	آ	0622

A.II.c. other options
Note that this script maps Arabic diacritics to Urdu diacritics, and also maps FDF7 to a single character superscript for PBUH (peace be upon him). Other choices could be needed here, depending on the spelling used in the lexicon and the data.

Control characters, when taken as part of a word, can cause spelling variations that are difficult to detect on the surface. These include various directional formatting and spacing characters. The presence of control characters prevents words from being recognized properly, causing problems for POS tagging and NE tagging. Spelling normalization might be extended to remove control characters.

A.III. parallel text

Here we list things that appear to be either errors or idiosyncrasies of particular writers. Variations that seem to be characteristic of Pakistani English in general are noted in Section B, below.

A.III.a. sentence level problems

A.III.a.i. mis-alignment

Sentence alignment was often thrown off by headlines or bylines that were not included on both sides. For example, the single English sentence below was split into three lines in the Urdu text.

JANG_ENG_20060331.86682
Karachi… staff reporter… the Intel Corporation has announced the results of world competition of video photography in its movie technology digital camera, and in the winners, Pakistan has also achieved success.

...کراچی
...رپورٹر اسٹاف
اعلان کا نتائج کے مقابلے عالمی کے فوٹوگرافی وڈیو میں کیمرے ڈیجیٹل ٹیکنالوجی مووی اپنی نے کارپوریشن انٹیل
ہے۔ کی حاصل کامیابی بھی نے پاکستان میں جن ہے کردیا

Sentence alignment was often disrupted by punctuation in web addresses and decimal numbers. Here we see a web address, www.lgsindh.pk, that has been split into three sentences. (note: in the third line, the word processor has displaced the first word, “pk”, to the left.)

JANG_URD_20060331.77486
وسیم اختر کی ہدایت پر دونوں کمیٹیوں کی سفارشات اور تجاویز عوام کی آگاہی کیلئے ویب سائٹ پر جاری
www.کردی گئی ہیں جو
lgdsindh.
پر دیکھی جاسکتی ہیں۔ pk

In this example we can see the effect of decimal points being mistaken for sentence final punctuation. Here, 6 lines become 12 lines. (note: in the Urdu section, leading numerals are displaced to the left by the word processor.)

JANG_ENG_20060331.88504
The ratio of GDP is expected to be between 6 and 6.6.
The imports would reach to $ 27.1 billion.
There is the trade lose of $10.4 billion.
The financial assets will grow up to 13.1 per cent.
The stocks of exchange of money become less by $ 1.29 billion.
The ratio of tax collection shrank to 18.6 percent from 30.8 percent.

JANG_URD_20060331.88504
6. سے6 شرح کی پی ڈی جی
ہے۔ امکان کا رہنے تک فیصد6
27. درآمدات
گی۔ جائیں پہنچ تک ڈالر ارب 1
ہے۔ سامنا کا خسارے تجارتی کے ڈالر ارب4
13. کر بڑھ اثاثے مالیاتی
گے۔ جائیں ہو فیصد1

1.ذخائر کے زرمبادلہ
گئے- ہو کم ڈالر ارب29
30. شرح کی وصولی ٹیکس
18.کر ہو کم سے فیصد8
رہی- تک فیصد6

We used word alignment probability as a way to detect poor sentence alignments. The worst sentences in terms of word alignment are derived from in the following files:

EMIL_URD_20051202.w-health-nation
SUK_URD_20060309.hist
EMIL_URD_20051202.w-legal-training
EMIL_URD_20051202.w-social-noise
JANG_URD_20060331.77094
EMIL_URD_20051202.w-health-nhs
JANG_URD_20060331.73492
JANG_URD_20060331.89463
EMIL_URD_20051202.w-legal-permit
EMIL_URD_20051202.w-health-warm

We also see sentence alignment problems with paragraphs that (apparently) derive from tables on web sites: The headings are placed differently within the derived Urdu and English text. See, for example, SUK_URD_20060309.hist. Here is the original table, as given in Monlingual_Text/Train/original_docs/SUK_URD_20060309.hist.html:

سکھر ڈسٹرکٹ میں چار تحصیلیں ھیں، جن کے نام، رقعبہ

آبادی درج ذیل ھیں-:

آبادي	رقعبہ	تحصیل	نمبر
374,178	274	سکھر	01
224,362	1319	روھڑی	02
64,646	2339	صالح پٹ	03
245,187	1233	پنو عاقل	04
9,08,373	5165		ٹوٹل

We found an English-language website with matching information, reading as follows:

There are 04 tehsils in file district. The area and population of each are as under:-

S.NO TEHSIL AREA IN Sq. M POPULATION

01	Sukkur	274	374,178
02	Rohri	1319	224,362
03	Saleh Pat	2339	64,646
04	Pano Akil	1233	245,187
Total		5165	9,08,373

Here is the way this table shows up in the derived text in English and Urdu. The English file lacks the numerical values altogether, while the Urdu file omits some of the data:

SUK_ENG_20060309.hist
There are 04 tehsils in file district.
The area and   population of each are as under:-
S.NNO
TEHSIL
AREA IN Sq.
POPULATION
Sukkur
Rohri
Saleh Pat
Pano Akil
Total

SUK_URD_20060309.hist
(Note: I have added translations in square brackets.)

سکھر [Sukkur]
چار میں ڈسٹرکٹ [district in four]
رقعبہ نام، کے جن هیں، تحصیلیں [tehsils UNK, body of name, area]
آبادی [population]
هیں ذیل درج :- [enter addenda UNK]
نمبر [number]
تحصیل [tehsil]
رقعبہ [area]
آبادي [population]
سکھر [Sukkur]
374,178
روهڑی [Rohri]
1319
224,362
پٹ صالح [Saleh Pat]
2339
64,646
عاقل پنو [Pano Akil]
1233
245,187

ٹوٹل [total]
5165
9,08,373

A.III.a.ii. leftover HTML

There appear to be bits of HTML tags left in some sentences; other sentences consist entirely of this material, which then leads to poor sentence alignment. Examples:

BBC_ENG_20051202.lhr_4thday_rza
Stephen Harmison/SEG>

JANG_ENG_20060331.79416
/> Saudi Arab: enquiry started against 46 Umarah companies HL>/>

JANG_ENG_20060331.74426
/>HL>

We also noticed the inclusion of sentence id numbers in some sentences. 8 English sentences include an SAP id number:

SAP20040321000017 Rawalpindi Nawa - i - Waqt in Urdu 20 Mar 04 p 14
SAP20040718000006 Rawalpindi Nawa - i - Waqt in Urdu 17 Jul 04 p 2
SAP20040807000043 Rawalpindi Nawa - i - Waqt in Urdu 06 Aug 04 p 10
SAP20041019000059 Rawalpindi Nawa - i - Waqt in Urdu 18 Oct 04 p - 5
SAP20041026000046 Rawalpindi Nawa - i - Waqt in Urdu 25 Oct 04 p - 2
SAP20050315000046 Rawalpindi Nawa - i - Waqt in Urdu 15 Mar 05 p 14
SAP20051022005002 Rawalpindi Nawa - e Waqt in Urdu 20 Oct 05 p - 3
SAP20051203002001 (Internet) Jamaatud Daawa Pakistan WWW - Text in Urdu 01 Dec 05

1 Urdu sentence consists solely of its APW id number:

APW_ENG_20030311.0775

A.III.b. word segmentation

A.III.b.i. run together words in the English text

We noticed that some ltf files in Parallel_Text/Train/Found/English/ltf have words run together. For example,

SAP_ENG_20050106.000079
The United States and the European countries didnot demonstrate the same spirit and enthusiasm for helping the tsunamivictims as they did to deal with natural catastrophes taking place inthe West.
....
Formal messages of "sympathy" have come from theUnited States and Europe, and some funds have also been announced, butthis process has not moved forward beyond verbaldeclarations.

We identified 21 files with significant numbers of run together words, using the spell checking program, Aspell, to find unknown words that could be split into two known words. While this process sometimes incorrectly attempts to split place names and acronyms, we considered suspicious any file that had more than 5 potential splits. These files are listed below.

SAP_ENG_20050106.000079	59 potential splits
SAP_ENG_20050110.000109	83 potential splits
SAP_ENG_20050119.000088	86 potential splits
SAP_ENG_20050125.000056	88 potential splits
SAP_ENG_20050128.000040	74 potential splits
SAP_ENG_20050204.000041	102 potential splits
SAP_ENG_20050225.000042	88 potential splits
SAP_ENG_20050225.000049	74 potential splits
SAP_ENG_20050227.000005	72 potential splits
SAP_ENG_20050228.000030	76 potential splits
SAP_ENG_20050303.000048	71 potential splits
SAP_ENG_20050303.000105	121 potential splits
SAP_ENG_20050307.000039	88 potential splits
SAP_ENG_20050308.000039	72 potential splits
SAP_ENG_20050311.000030	78 potential splits
SAP_ENG_20050314.000103	70 potential splits
SAP_ENG_20050317.000048	97 potential splits
SAP_ENG_20050319.000019	61 potential splits
SAP_ENG_20050322.000094	120 potential splits
SAP_ENG_20050324.000059	59 potential splits
SAP_ENG_20060107.005002	9 potential splits

Aspell identified a total of 1648 potential splits in these files; some of these are errors, as when Aspell tries to split place names, but most of them are necessary corrections.

There were also 2 files in which words were interrupted with other words, LON_ENG_20050601 and LON_ENG_20050801. For example,

LON_ENG_20050601
it's estiWrite to mated that at least Talking 150,000 people have quit Point - since the ban was introsee p9 duced in March 2003.

which should presumably read,
it's estimated that at least 150,000 people have quit since the ban was introduced in March 2003. Write to Talking Point - see p9

A.III.b.ii. spacing errors

Spacing issues are a general characteristic of Urdu writing, and not a problem specific to the creation of the LCTL Urdu parallel text. Because spacing causes problems for word alignment and named entity tagging, we include it here.

Urdu speakers do not necessarily treat space as a tokenization marker, rather, they just want the script to appear visually correct. When characters are non-joiners, the appearance of space is present, and so two words may be typed without a space character between them. Similarly, a space may be tolerated within a single word, if it does not disrupt any links between joining characters. (See references, Section D.)

For example,
JANG_ENG_20060331.76762
Dubai Jang News Al Qaeda's leader and one of Osama Bin Laden's deputies, Ayman Al Zawahiri, has appealed to Muslims all over the world to help victims of the earthquake in Pakistan.
پاکستانی وہ کہ ہے کی اپیل کو مسلمہ امت نے الزواہری ایمن نائب کے لادن بن اسامہ اور رہنما کے نیوزالقاعدہ دبئیجنگ
کرے۔ مدد ممکن ہر جلد کی متاثرین زلزلہ

Here we see that the words Dubai دبئی and Jang جنگ have been run together as دبئیجنگ , while the words News نیوز and Al Qaeda القاعدہ (one word in Urdu) have been run together as نیوزالقاعدہ (the effect is "DubaiJang NewsAlQaeda"). Visually, there is no problem separating the words News and Al Qaeda, since the characters are non-joiners; Dubai and Jang, however, should have been held distinct.

with spaces نیوز القاعدہ without spaces نیوزالقاعدہ
with spaces دبئی جنگ without spaces دبئیجنگ

In addition to the general spacing issue, we note that in Urdu postpositions and conjunctions are commonly written attached to an adjoining word. We see this in the parallel text, and in the laf files that contain the tagged named entities. In this example, we find "andBrazil" اوربرازیل:

WSJ_ENG_20060331.0515
Small crops are grown in Pakistan, France, Spain, Italy, Belgium and Brazil, but their quality can't compare to that of Indian psyllium
بھارتی لیکن ہے جاتی کی کاشت میں اوربرازیل بیلجیئم اٹلی، اسپین، فرانس، پاکستان، پر پیمانے چھوٹے کی بیج اس
ہے۔ حاصل فوقیت پر ان کو اسبگول

This instance of "andBrazil" also shows up in the named entity annotations, tagged as type "LOC":

```
WSJ_URD_20060331.0515.laf.xml
   <EXTENT>اوربرازیل</EXTENT>
```

Looking through the NE annotations, we find other examples with postpositions and conjunctions added to the beginning or end of nouns, such as:

نےاسرائیل	ne + Israel
نےامریکی	ne + America
بلیئراور	Blair + aor

A.III.c. word level errors

A.III.c.i. English spelling errors

There are typographical errors (e.g., filds for fields, iinto for into), and also some very confused spellings, such as “Serket Koat” for "Circuit Court" and “Querkistan” for (presumably) "Kyrgyzstan". There are some subtler errors, like drought spelled draught (the British spelling of draft).

Names in particular may have many variant spellings. Sometimes these are legitimate variants. With public figures, we are able to determine a correct spelling; other times we cannot tell. For an example of the wide range of spellings given in the English translations, here are the variations of the first and last name of a cricket player, Marcus Trescothick:

Marcus	Trescothick
Marcus	Trescothick
Marcos	Trescothik
Mark	Trescothek
Marks	Treskothic
Markus	Treskothiek
Marques	Threstock
	Thresthock
	Trescothech
	Triscothick

Susbstitution errors, like draught for drought, have to be evaluated in context, but many misspellings can be automatically detected. We have noted more than 700 unique words that are clearly misspelled.

Some words flagged by our spell checker that appear with enough frequency that we consider they may be normal in Pakistani English, such as increasement meaning increase, and juditional meaning judicial. We give more information on these in Section B.VII.b.

Some of the most noticeable spelling errors come from sections of our derived text for which we cannot currently identify the source files. The Urdu portions can be found in the Monolingual text section. Two such sections are:
BBC_URD_20050419.pope_election_ra
BBC_URD_20040628.michael_moore_911_an

From the first section, we find the following English spelling errors:

written	meaning
Pop John Paul	Pope John Paul
Sent Peter Becelica	St Peter’s Basilica
Bandickt	Benedict
Retzinger	Ratzinger

cordinal	cardinals

From the second section, we find this:

written	meaning
Foreign Byte 9/11	Fahrenheit 9/11

A.III.c.ii. grammar errors

There is some non-idiomatic English, as in this sequence:

BBC_ENG_20051208.audiovideo_quake
It was written on a banner 'good new'.
The picture of earthquake and the condition in video CD now.
This banner doesn't seen again.
One shopkeeper told that the member of any organisation took it droving.

Our native speaker consultant has also noticed some translations which reverse the meaning -- these include word-choice errors as well as sentence structure discrepancies. In the first example, the word negative منفی in the Urdu text corresponds to the word positive in the English text:

JANG_ENG_20060331.89153
The helicopter crushed by any technical reason means while the Iraqi army; have arrested 70 people in Baghdad, there for the Iraqi election commission have announced that the positive result of election will come, after two week
. گا لگے وقت کا ہفتے دو مزید ابھی میں آنے نتائج منفی کے انتخابات کہ ہے کہا نے کمیشن الیکشن عراقی اثنأ دریں
literally, "therefore (darian asna) Iraqi Election Commision ergative said is for election genitive negative results coming in yet additional two week genitive hour [lge] will"

In this WSJ article (presumably translated from English to Urdu), our consultant says the phrase "one-third" is modifying the wrong part of the sentence. The Urdu sentence states that the projected Afrikaner homeland is to cover 1/3 of South Africa territory.

WSJ_ENG_20060331.1760
But their ideal of an Afrikaner homeland, an all-white reserve to be carved out of present-day South Africa, is a mainstream desire of the right-wing, which embraces about one-third of the country's five million whites.
کی بازو دائیں خیال کا کرنے قائم وطن افریکانر کے فاموں سفید صرف اور صرف اندر کے افریقہ جنوبی موجودہ لیکن
ہے۔ کرتی احاطہ کا آبادی تہائ ایک کے فاموں سفید ملین پانچ کے ملک جو ہے خواہش اہم

A.III.c.iii. additional information in the translations

Some of the English translations contain explanations within the translations. These can be translations of borrowed words, or notations like “previous three words in English”. While not exactly an error, this can be problematic for word alignment.

The first example here contains a comment about a name, and a note pointing out a transliterated English word (سکرپٹ /skrpt./ “script”). The second example has a notation about English words, which appear in both transliteration and in English in the Urdu text (stick سٹك and carrot کرٹ).

SAP_ENG_20041019.000107
Right from Amrinder Singh [Indian Punjab chief minister] to Gujral, all Indian leaders are reading the same script [preceding one word in English] that has been given to them by the Indian government.

امریندرسنگھ سے گجرال تک سب وہی سکرپٹ پڑھ رہے ہیں جو بھارتی حکومت نے انہیں یاد کرایا ہے۔

SAP_ENG_20050830.000140
US Secretary of States Condoleezza Rice, while testifying before the Senate Foreign Relations Committee, called it a policy of carrots and sticks [preceding three words in English].

کنڈولیزا رائس نے سینٹ کی امور خارجہ کمیٹی کے سامنے بیان دیتے ہوئے اسے سٹک (STICK) اور کرٹ (CARROT) یعنی چھڑی اور گاجر پالیسی کا نام دیا۔

A.III.d. numerals and punctuation

A.III.d.i. number reversals

We have noticed reversed numbers in some sentences, especially with decimals, also with some years. We have checked the unicode codepoints and determined that this is not an effect of the word processor display. Could this be due to confusion by Urdu writers when attempting to enter numbers for correct left to right display?

In the first example, the Urdu reads “01 pound” instead of “10 pound”. In the second example, we find “57%” instead of “75%”.

EGO_ENG_20060222
They cost more than regular light bulbs (starting at $5), but can use 75% less electricity and last years longer.

ںیہ ےترکاوراس یک ل یماع موبلب ںےس $5(ںیہ ےتوہ ےگنہم ےس رشوع یلجب مک 57% نکیل)ںیہ ےتوہ اامعتسل
ںیہ ےتلچ کت

NACC_ENG_20060222.ue
The renewal subscription is currently £10 per year.

.ےہ لاس یف ڈنوپ 01 ہدنچ دیدجت لاحلا یف

More examples of numeral reversal can be found in LON_ENG_20050601.

A.III.d.ii. creative use of Roman numerals

Some of the English translations use Roman numerals to spell out ordinal numerals.

BBC_ENG_20051112.1test_1stday

He is the VIth aged test captain cricketer of after the IInd world war.

BBC_ENG_20051118.sibtain_an
My sister was also study in IInd class.

BBC_ENG_20051201.lhr_3rdday_rza
Before this England team scored 288runs in its Ist innings.

A.III.d.iii. creative use of punctuation

In punctuating the Urdu text, the writers sometimes use Urdu punctuation, and sometimes use English punctuation. There is variation in the punctuation of numerals, which sometimes have a comma for a decimal point, may or may not have a slash before a unit word, and may or may not use a hamza as a year marker.

The tatweel, 0640, designed to lengthen characters for better visual presentation, appears in isolation as a short horizontal mark; this is sometimes used in the data as a substitute for the Urdu full stop character 06D4, or in multiples as a kind of hyphen. For examples, see:

MEHR_URD_20060331.148998
WSJ_URD_20060331.1760
WSJ_URD_20060331.1121

There is also some creative use of punctuation in the English text:

In JANG_ENG_20060331.90088, we find the lowercase letter o used in place of a degree mark: “oC” meaning “degrees Centigrade”.

There is one example of an emoticon:
MAR_ENG_20050127.0716
You haven't missed much :).

A.IV. Part of Speech Tagger

The POS tagger works best if the spelling is first normalized (Arabic forms > Urdu forms). The tagger also tags certain control characters (e.g., feff), so it is better to remove these first.

It would be useful to be able to tag words within sentences. To get the correct input format for the tagger, we had to change sentences into wordlists and append dummy tags.

The tag set differs from the Penn or MXPOST tagsets.

Inconsistency with punctuation: Punctuation marks should be tagged as PUNCT, but in our test files, we saw punctuation tagged as OTHER, or, in some sentences, final punctuation was tagged with that same punctuation mark as the tag value.

A.V. Named Entity Tagger

We noticed that the named entity tagger skips the last line of the input file.

We found it necessary to split up large files when using the tagger.

We note in Section A.III.b.ii. that the elements tagged as named entities in the laf files may contain extraneous material, such as adjoined postpositions or conjunctions. Tagged phrases sometimes include words that do not logically belong to the named entity. We also see tagged named entities with attached punctuation, and sometimes, punctuation alone tagged as a named entity (for example, an single-character ellipsis in laf file JANG_URD_20060331.77336, a three-character ellipsis in JANG_URD_20060331.77370).

Postpositions themselves are often tagged as named entities, and may occur with a range of type designations. The ergative postposition /ne/ is tagged variously as PER, LOC, and ORG, for example, while the genitive postposition /ke/ can be PER, LOC, ORG, or TIME.

A.VI. transliterator

The transliterator considers all vowel permutations between each pair of consonants. This makes it take too long or fail on large files, and introduces implausible translations (e.g., NAWAZ for نیوز/nywz/ "news").

There are spelling errors in English word frequency list "governmnt", "industrys", "indpendent"; however, because these are lower in frequency than the correctly spelled words, this might not be a problem. (For example, “government” has its frequency listed at 1,655,352, while the misspelled “governmnt” has a listed frequency of only 7.)

The transliterator would be more useful if it could be applied to tagged words within a sentence (for example, we would like to tag a sentence for words not found in the lexicon, and then apply transliteration to just those words).

Positive: Since the transliterator uses a word frequency list, it would be possible to create different word lists for different domains. This might be helpful in providing some context-sensitive translation. For example, one might maintain a different word list for names as opposed to common nouns, or for newswire as opposed to travel utterances.

B. language-specific issues

Here we discuss general characteristics of Urdu or of Pakistani English that cause difficulties for the machine translation. We list them separately from Section A, which deals with problems specific to the LCTL materials.

B.I. Spelling
B.II. Spacing
We have already discussed spelling normalization (Sections A.I.i. and A.II.) and spacing issues (Section A.III.b.ii.). While these are general problems affecting Urdu text processing, they have a specific impact on the use of the LDC materials, as explained in those sections.

B. III. Unicode

B.III.a. Medial [e]

There is a generally known issue with Unicode support of Urdu for the spelling of a medial [e] 06D2. This letter must be written [i] 06CC in order to display properly in the middle of a word.

B.III.b. Right to left formatting

In order to create the appropriate directional formatting (right to left for most text, with embedded left to right numerals and Latin characters), a number of Unicode control characters are used. These can cause problems since a word with an attached control character is computationally distinct from the same word without the control character.

B.IV. Names and homophones

Many Urdu names are homophonous with other Urdu words: Aziz = beloved; Shoukat = dignity; Malik = state; Mir = chief; etc. This can be problematic if we are attempting to pre-translate named entities, for example.

B.V. Acronyms

Translated acronyms like BBC are generally written out with one word per letter (bi bi si). This creates a problem for word alignment; there can also be confusion when the elements of the acronym coincide with existing Urdu words. For example, a word-by-word translation of “bi bi si” is “teal teal thirty”.

We have noticed two acronyms in the data that follow the English pattern of one letter per word: the Pakistani agencies NEPRA (National Electric Power Regulatory Authority) and WAPDA (Water and Power Development Authority). These are written as single Urdu words: /nypra/ نیپرا and /wapd.a/ واپڈا

The honorific ﷺ PBUH, may be translated into English as a phrase or an acronym:

(May God bless him)
(Peace be upon him)
(PBUH)
(pbuh)

B.VI. Borrowing

B.VI.a.Prevalence of borrowing

Urdu writers may borrow from English even when native words are available. so sifar 'zero' is less frequent than its borrowed counterpart, 'zyro', and we see the numeral 9 sounded out as in English, ‘nine’ along with native 'nau'. Some further examples of borrowing:

JANG_ENG_20060331.86682

The chief operating officer of Intel movie Mr., Philip Moorish has said that these photographers have made new standards taking best photographs from digital cameras of Intel.

انٹیل مووی کے چیف آپریٹنگ آفیسر فلپ مورش نے کہا ہے کہ ان فوٹو گرافروں نے انٹیل کے ڈیجیٹل کیمرے سے
بہترین فوٹو گرافی کرتے ہوئے نئے معیار قائم کئے ہیں۔

word by word, with transliterations underlined:
<u>Intel</u> <u>movie</u> ke <u>chief</u> <u>operating</u> <u>officer</u> <u>Philip</u> <u>Moorish</u> ne kha he kh an <u>photo</u> <u>graph</u>-oN ne <u>Intel</u> ke <u>digital</u> <u>camera</u>-e se bhtryn <u>photo</u> <u>graphy</u> krte hoie nie mayar qaim kie hyN

JANG_ENG_20060331.73954
In the top ten, India comes in the last
اس سرٹیفکیشن کی ٹاپ ٹین میں بھارت آخری نمبر پر ہیں۔

word by word, with transliterations underlined:
as <u>certification</u> ky <u>top</u> <u>ten</u> myn bhart Ajry <u>number</u> pr byN

B.VI.b. Identifying Borrowed Words
Transliteration of borrowed words may help the machine translation process, but it is not always easy to distinguish borrowed words. Borrowing may create homophones with existing words, such as English “Bill” = Urdu “twist”, English C = Urdu “thirty”; English K = Urdu postposition ke; etc.

One clue to help identify borrowings: English alveolar d, t are further back than dental Urdu d, t, and so are usually written in Urdu with retroflex t ٹ and d ڈ . Seeing these characters is a clue that the word may be borrowed.

B.VI.c. interaction of borrowed words and native morphology:
Urdu speakers may borrow an English word, and then apply an Urdu inflection. So we get Urdu plurals in leader-oN, smuggler-oN, etc. Here, we see the borrowed word “wicket” with the Urdu plural suffix: وکٹیں

JANG_ENG_20060331.86279.ltf.xml
Imran Farhat of HBL took three wickets for 18 runs, and Abdul Rahman took 5 wickets for 120 runs.
ایچ بی ایل کے عمران فرحت نے 18/ رنز دے کر تین وکٹیں اور عبدالرحمن نے 120/ رنز دے کر پانچ کھلاڑیوں کو
آؤٹ کیا۔

This blended morphology makes it more difficult to identify and transliterate borrowed English words.

B.VI.d. Borrowing complicates numeral phrase patterns:
The numeral pattern is complicated by the option of using transliterated English unit words like ملین /myln/ “million”. Since the unit words divide at different quantities for Urdu and English, by choosing one system or the other, the Urdu speaker may have multiple possible ways to express a single quantity. Here, لاکھ /lakh/ is the Urdu word for 100,000.

English	3.5 million
Urdu	3.5 myln
Urdu	35 lakh

Note: The Urdu language also has words to express "plus one half" ساڑھے /sarhe/, "one and a half" ڈیڑھ /dyrh/, and "two and a half" ڈھائی /dha'y/, and these create additional possibilities. For example,

English: 250,000
Urdu: dha'y lakh (2½ 100,000)

English: 5:30
Urdu: sarhe panch bje (+½ 5 o'clock)

B.VII. Pakistani Vocabulary

B.VII.a. British vs American English

There is of course a British character to Pakistani English, and this affects machine translation. In looking at the parallel text, we find that the English translations include both British and American spellings, with slightly more American spellings. Computationally, words like "honor" and "honour" are distinct, so it is useful to normalize to either British or American spelling.

B.VII.b. Pakistani English: local vocabulary

We see specific local words like godowns and incharge; there are also certain abbreviations like s/o for "son of". We also see differences in the use of plural forms (machineries instead of machinery).

We see vocabulary choices that are atypical in American English (expulsive for explosive), as well as neologisms like juditional for judicial. It is not clear whether all of these are expressions of the local English dialect, or whether some of them are translation errors.

Examples of local vocabulary:

Found	Meaning
godowns	warehouses
tubewells	simple wells
nazim	a local official
incharge	a job title
machineries	machinery
informations	information
furiousness	fury
satisfactable	satisfactory
expulsive	explosive
redressal	redress

Found	Meaning
increscent	increase
incensement	increase
increasement	increase
pilgrimers	pilgrims

juditional	judicial	
documental	documentary	
picturised	photographic	
s/o	son of	
w/o	wife of	
ODI	one day international	(cricket match)

C. SCREAM Lab Innovations

C.I. Lexicon
C.I.a. Corrections and Additions

C.I.a.i. Additions:
We made 2252 additions to the LCTL Urdu lexicon to create our augmented lexicon, “auglex”. Because of other corrections (see below), we also removed or collapsed 671 entries, giving us a final auglex with 1581 more entries than the original LCTL lexicon.

auglex dictionary	
retained LCTL entries	25,692
auglex additions	2,252
total	27,944

In adding words to the lexicon, we focused on words occurring with high frequency in the training data. We collected the highest frequency words that were not found in the lexicon, and prioritized these for translation, either by examination of the word in context, or through consultation with a native speaker of Urdu. We also focused attention on words found in headlines, unknown words tagged as named entities, and words that remained persistent unknowns in the output of our machine translation process. Other words were added opportunistically as we translated particular sentences in order to analyze alignment or spelling problems.

We experimented with dividing the auglex into common words and named entities, for the purpose of pre-translating named entities. We further divided the named entity section into names with distinct meanings, and names with multiple English meanings/spellings, as we do not want to pre-translate a word as a name if it has other possible meanings (e.g., malik = state). There are 2039 entries in the regular NE sub-lexicon, and 171 entries in the multiple-NE sub-lexicon.

C.I.a.ii. Corrections:
We do not have a current count of the number of corrected entries. We made systematic corrections for reversed phrases and comma-separated glosses. In addition, we corrected some spelling errors and POS errors as we came across them, and we re-ordered some gloss entries to make the more common meaning the first meaning. At the same time, we normalized spelling in the augmented lexicon. These spelling changes caused us to collapse many entries that differed only in the use of diacritics. Altogether, we removed 671 entries from the original LCTL lexicon.

Here is an example of entries that were collapsed. The original lexicon has:

```
<ENTRY id="LEX-URD-01932100">
  <WORD>اَتِھر</WORD>
  <POS>NONE</POS><GLOSS source="urdu.lex">changeable</GLOSS>
  <POS>NONE</POS><GLOSS source="urdu.lex">inconstant</GLOSS>
</ENTRY>

<ENTRY id="LEX-URD-01105400">
  <WORD>اتھر</WORD>
  <POS>NONE</POS><GLOSS source="urdu.lex">changefull</GLOSS>
  <POS>NONE</POS><GLOSS source="urdu.lex">lubric</GLOSS>
  <POS>NONE</POS><GLOSS source="urdu.lex">lubrical</GLOSS>
  <POS>NONE</POS><GLOSS source="urdu.lex">restless</GLOSS>
  <POS>NONE</POS><GLOSS source="urdu.lex">unsteady</GLOSS>
</ENTRY>
```

But our derived auglex has no entry for LEX-URD-01932100 اَتِھر . After removal of the diacritic,LEX-URD-01932100 combines with LEX-URD-01105400:

```
<ENTRY id="LEX-URD-01105400">
  <WORD>اتھر</WORD>
  <POS>NONE</POS><GLOSS source="urdu.lex">changefull</GLOSS>
  <POS>NONE</POS><GLOSS source="urdu.lex">lubric</GLOSS>
  <POS>NONE</POS><GLOSS source="urdu.lex">lubrical</GLOSS>
  <POS>NONE</POS><GLOSS source="urdu.lex">restless</GLOSS>
  <POS>NONE</POS><GLOSS source="urdu.lex">unsteady</GLOSS>
  <POS>NONE</POS><GLOSS source="urdu.lex">changeable</GLOSS>
  <POS>NONE</POS><GLOSS source="urdu.lex">inconstant</GLOSS>
</ENTRY>
```

We noted in Section A.I.g.i. that the lexicon contains spelling errors in the English glosses. We tried using the Aspell spellchecker program to correct these errors. We had the program consult its references for both British and American English, and we set up the program to ignore uppercase words (to protect named entities), hyphenated words, and numbers. Even with these conditions, the results were not viable: While Aspell successfully corrects some misspelled words, the lexicon also contains rare words that are unknown to Aspell, and these are changed to more common words.

Lexicon Gloss Entry	Aspell Suggestion
sataition	satiation
dialtory	dilatory
plutonic [correct]	Platonic
pemphigus [correct]	Memphis

Also problematic were the words that reflect the creative use of English morphology by Pakistani speakers (see Section B.VII.b). Spelling correction is inappropriate here.

Lexicon Gloss Entry	Meaning	Aspell Suggestion
reposal	repose	reprisal
needments	necessities	tenements

Spelling correction of the English gloss entries will have to be a supervised process.

C.I.b. Inflections
In addition to defining the 2,252 high frequency words described above, we added inflected forms to the lexicon by bringing in the forms of the Humayoun morphological analyzer. The Humayoun morphological analyzer contains 4,045 Urdu stems, with 106,931 inflected forms, but lacks English definitions. We brought these into the lexicon, combining them with any existing gloss and morph entries. The resulting "auglex-plus-humayoun" lexicon has 44,290 entries.

auglex-plus-humayoun dictionary	
retained LCTL entries	25,692
auglex additions	2,252
Humayoun additions	16,346
total	44,290

C.II. Spelling Normalization
Building on the LDC provided script, UrduPresentation2Regular.tcl, we developed a spelling normalization program that converts from Arabic script to Urdu script. This program includes the corrections noted here in Section A.II, as well as mappings for the Arabic presentation forms. Our partners at MIT created a different normalization program that converts from Urdu script to Arabic script.

C.III. Parallel Text
Note: most of our work with the parallel text was done subsequent to a lowercasing process, so some proper names are cited here in lower case.

British vs. American English
We used the Aspell spell checker to count the number of British vs. American spellings; finding a slight majority for American spellings, we then used the Varcon program to normalize the text to American spellings.

Spelling Errors: English
Many misspellings can be automatically corrected. We have compiled a list of over 700 words that are clearly misspelled, together with their corrected forms, for use in automatic spelling correction via the Varcon program. For example,

original	correction
tonigt	tonight
thusday	thursday

webside	website

We also created a list of 3800 protected words: These are words that we believe to be local names or vocabulary that should not be corrected (e.g., sindh), along with ordinary English words that aren't covered by the Aspell dictionary (e.g., biosciences) and website terms like bbcpersian.

Hand editing: English
For some spelling errors, we have had to hand edit files. This includes removing leftover HTML and examining words in context to determine whether we have a misspelled word or a correctly spelled named entity (e.g., the word gardener vs. the name gardner).

Split and Compounded Words: Urdu
In Section A.III.b.ii., we give examples of Urdu words in which conjunctions and postpositions have been attached to the word, as in the example بلیئراور blairaor = Blair + and. We have developed a program to automatically split off potential conjunctions and postpositions from unknown words. This corrects many of the run-together words, but also has the potential to split up legitimate unknown words like San Francisco سین فرانسسکو and Orlando اورلینڈو (taking –ko as a postposition, and aor- as the conjunction, respectively).

Split and Compounded Words: Urdu and English
We have also experimented with a program that identifies split and compounded words by comparing bigram and unigram counts. This is a language-independent technique, that can find errors in the English as well as the Urdu.

C.IV. POS tagging

When using the LCTL POS tagger, we first normalize our data to Urdu spelling and remove control characters; then we apply a program separates each sentence into one word per line and adds a dummy tag, in order to provide the correct input to the tagger. After tagging, we apply another program to reconstitute the file into running text. We also apply a program to convert the LCTL POS tags to either the MXPOST or the Penn tagset.

C.V. NE tagging

When using the LCTL NE tagger, we first split large files into subfiles, adding a dummy line to the end of each file (since the tagger skips the last line). After tagging, we reconstitute the files.

C.VI. Transliteration

We created a new transliterator program that improves performance with words that have been borrowed from English into Urdu. We followed the rule-and-dictionary approach of the LCTL transliterator, but we mapped Urdu characters to sounds, and used the CMU English pronunciation dictionary instead of a word frequency list in English spelling. As in the LCTL transliterator, we first matched consonants, and then dealt with vowels, but instead of considering every vowel permutation between every consonant pair, we made a language-specific mapping. We mapped Urdu vowel characters to ipa sounds, and enabled the program to skip lax vowels in the English entries. We reduce the search space by only considering English words that match the consonantal pattern during the vowel mapping phase, making the program more efficient.

C.VII. Stemming

We have implemented a stemming program to help us make better use of the lexicon for inflected forms, using Ramanathan and Rao, A Lightweight Stemmer for Hindi, as our prototype. Our stemmer removes the longest possible inflection from the end of a word, maintaining a stem of at least two characters. When POS information is available, the program distinguishes nominal, verbal, and adjectival inflections. We use the stemmer to create factored data. The output of the stemmer can also be used to look up words that match either a word or stem entry in the LCTL lexicon.

Some restricted stemming is also useful in transliteration, to handle cases of Urdu morphology applied to borrowed English words (see Section B.VI.c. for examples).

D. References

Humayoun, Urdu morphological analyzer http://www.lama.univ-savoie.fr/~humayoun/UrduMorph/

Ramanathan and Rao, A Lightweight Stemmer for Hindi
http://computing.open.ac.uk/Sites/EACLSouthAsia/Papers/p6-Ramanathan.pdf

Atkinson, Kevin. VARCON (Variant Conversion). http://wordlist.sourceforge.net/varcon-readme

Aspell: http://aspell.net/

information on Urdu spacing issues:
http://www.crulp.org/Publication/papers/2007/spelling_error_trends_in_urdu.pdf
http://www.lc-star.org/pccdocs/Pakistan_Urdu_LSP_V1.0.pdf

LIST OF ACRONMYS & GLOSSARY

711 HPW	711th Human Performance Wing
A3Metric	A collection of SCREAM lab developed software tools to examine and modify results of word alignment
A3PartitionFilter	A software tool to include out-of-domain data when performing word alignment
AFLAT	African Language Technology
AFRL	Air Force Research Laboratory
AM	Acoustic Model
Aspell	a free open source spell checker
ASR	automatic speech recognition
BLAS	Basic Linear Algebra Subroutine
BLEU	Bilingual Evaluation Understudy, an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another[24].
BTEC	Basic Travel Expression Corpus
CB	Callison-Burch
Centrify Express	software that provides Linux integration into Microsoft Active Directory
Ceph	a unified, distributed storage system
CER	Character Error Rate
CMLLR	Constrained Maximum Likelihood Linear Regression
CPAN	Comprehensive Perl Archive Network
CPU	Central Processing Unit
CSLM	Continuous Space Language Model
CSR	Continuous Speech Recognition
DoD	Department of Defense
DTIC	Defense Technical Information Center
FFMPEG	FFmpeg is a free software / open source project that produces libraries and programs for handling multimedia data.
Fisher	an English corpus of conversational telephone speech
FMCF	Foreign Media Collaboration Framework
GB	10^9 bytes
GF	Grammatical Framework
Gigaword	a text corpus produced by the Linguistic Data Consortium
GIZA	Training program that learns statistical translation models from bilingual corpora, part of The EGYPT Statistical Machine Translation Toolkit.
GIZA++	GIZA++ is an extension of the GIZA program.
GLOSS	Global Language Online Support System
GMM	Gaussian Mixture Model
GPGPU	General-purpose graphics processing unit
GUI	Graphical User Interface
Hadoop	an open source software framework for data-intensive distributed applications
HPW	Human Performance Wing

[24] http://en.wikipedia.org/Wiki/BLEU

HDecode	Cambridge University large vocabulary continuous speech recognizer
HLDA	Heteroscedastic Linear Discriminate Analysis
HLT	Human Language Technology
HMM	hidden markov model
HTK	hidden markov model toolkit
HTML	HyperText Markup Language
HTS	HMM-based Speech Synthesis System
HUB4	1997 broadcast news corpus
ICER	Information Operations Cyber Exploitation Research
ILR	Interagency Language Rating
IWSLT	International Workshop on Spoken Language Translation
Java	Java refers to a number of computer software products and specifications from Sun Microsystems that together provide a system for developing application software and deploying it in a cross-platform environment.
Joshua	an open source MT system developed at Johns Hopkins University
KTH	Royal Institute of Technology
LCTL	Less Commonly Taught Languages
LDC	Linguistic Data Consortium
Linux	A Unix-like free and open source Operating System.
LM	language model
LVCSR	Large Vocabulary Continuous Speech Recognizer
Metadata	Metadata is data providing information about one or more other pieces of data.
Meteor	an automatic MT evaluation metric
Meteor/NEXT	an automatic MT evaluation metric based on Meteor
MFCC	Mel-Frequency Cepstral Coefficient
MIT	Massachusetts Institute of Technology
MIT/LL	Massachusetts Institute of Technology Lincoln Laboratory
ML	Maximum Likelihood
Moses	a free software statistical MT engine
MPE	Minimum Phone Error
MT	Machine translation (MT) is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another.
MT Eval	Machine Translation Evaluation
MySQL	MySQL is a relational database management system (RDBMS) that runs as a server providing multi-user access to a number of databases.
NASIC	National Air and Space Intelligence Center
NFS	Network File System
Newscom	A provider of high-quality images, video and other supplemental content from a vast array of global sources.
NILE	A word alignment software package
NIST	National Institute of Standards and Technology
NLP	natural language processing
NLTK	Natural Language Toolkit (http://nltk.org)
OCR	Optical Character Recognition

OGS	Open Grid Scheduler, an open-source project based on SGE
OOV	Out Of Vocabulary
Parex	Paraphrase Extraction
PDF	Portable Document Format (Adobe)
Perl	Perl is a high-level, general-purpose, interpreted, dynamic programming language.
PHP	PHP: Hypertext Preprocessor is a widely used, general-purpose scripting language that was originally designed for web development to produce dynamic web pages.
PLP	Perceptual Linear Prediction
PostCAT	Posterior Constrained Alignment Toolkit
Python	Python is an interpreted, general-purpose high-level programming language whose design philosophy emphasizes code readability.
PLF	Python Lattice Format
RAID	RAID, an acronym for Redundant Array of Independent Disks is a technology that provides increased storage reliability through redundancy.
RDBMS	Relational DataBase Management System
SALT	Speech and Language Translation
SAT	Speaker Adaptive Training
SCREAM	Speech and Communication Research, Engineering, Analysis, and Modeling
SGE	Sun (now Oracle) Grid Engine, an open-source batch queuing system
SLF	Standard Lattice Format
SMT2	Statistical Machine Translation software created by MIT/LL and the SCREAM Laboratory that interfaces with Moses.
SLTS	Spoken Language Translation Systems
SOA	Service Oriented Architecture
SSO	Single Sign-On
SRILM	Stanford Research Institute Language Modeling toolkit
Sphinx-4	an open source large vocabulary continuous speech recognition engine
SS	Speech Synthesis
Systran	A commercially available MT software system
TB	10^{12} bytes.
TDT4	Topic Detection and Tracking corpus
TED Talks	TED (Technology, Entertainment, and Design) is a global set of conferences owned by the private non-profit Sapling Foundation, formed to disseminate “ideas worth spreading”[25] (http://www.ted.com/).
TextCat	A text language classifier
Thrax	an extractor for synchronous context-free grammars for use in MT
TRANSTAC	Spoken Language Communication and Translation System for Tactical Use
Treebank	A text corpus in which each sentence has been annotated with syntactic structure
TTS	Text Translation Systems
TW	Translator Workbench

[25] http://en.wikipedia.org/wiki/Tedtalks

URL	Uniform Resource Locator
VarCon	variant conversion
WER	Word Error Rate
WMT	Workshop on statistical Machine Translation
WMT11	2011 Workshop on statistical Machine Translation
WSDL	Web Services Description Language
WSJ	Wall Street Journal
XFS	a high-performance journaling file system created by Silicon Graphics
XML	eXensible Markup Language
ZeroMQ	a high-performance asynchronous messaging library